Article No. jmbi.1999.3444 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 296, 57–86

JMB

Fully Synthetic Human Combinatorial Antibody Libraries (HuCAL) Based on Modular Consensus Frameworks and CDRs Randomized with Trinucleotides

Achim Knappik[1]*, Liming Ge[1], Annemarie Honegger[2], Peter Pack[1] Melanie Fischer[1], Günter Wellnhofer[1], Adolf Hoess[1], Joachim Wölle[1] Andreas Plückthun[2] and Bernhard Virnekäs[1]

[1]*MorphoSys AG, Lena-Christ-Str. 48, 82152, Martinsried/ Munich, Germany*

[2]*Biochemisches Institut Universität Zürich Winterthurerstrasse 190 CH-8057, Zürich Switzerland*

*Corresponding author

By analyzing the human antibody repertoire in terms of structure, amino acid sequence diversity and germline usage, we found that seven V_H and seven V_L (four Vκ and three Vλ) germline families cover more than 95% of the human antibody diversity used. A consensus sequence was derived for each family and optimized for expression in *Escherichia coli*. In order to make all six complementarity determining regions (CDRs) accessible for diversification, the synthetic genes were designed to be modular and mutually compatible by introducing unique restriction endonuclease sites flanking the CDRs. Molecular modeling verified that all canonical classes were present. We could show that all master genes are expressed as soluble proteins in the periplasm of *E. coli*. A first set of antibody phage display libraries totalling 2×10^9 members was created after cloning the genes in all 49 combinations into a phagemid vector, itself devoid of the restriction sites in question. Diversity was created by replacing the V_H and V_L CDR3 regions of the master genes by CDR3 library cassettes, generated from mixed trinucleotides and biased towards natural human antibody CDR3 sequences. The sequencing of 257 members of the unselected libraries indicated that the frequency of correct and thus potentially functional sequences was 61%. Selection experiments against many antigens yielded a diverse set of binders with high affinities. Due to the modular design of all master genes, either single binders or even pools of binders can now be rapidly optimized without knowledge of the particular sequence, using pre-built CDR cassette libraries. The small number of 49 master genes will allow future improvements to be incorporated quickly, and the separation of the frameworks may help in analyzing why nature has evolved these distinct subfamilies of antibody germline genes.

Keywords: human antibodies; phage display library; combinatorial immunoglobulin repertoire; antibody expression; trinucleotide mutagenesis

Introduction

The selection of antibody fragments from libraries using enrichment technologies such as phage-display (Smith & Scott, 1993), ribosome display (Hanes & Plückthun, 1997), bacterial display (Georgiou *et al.*, 1997) or yeast display (Kieke *et al.*, 1997) has proven to be a successful alternative to classical hybridoma technology (for recent reviews,

Present addresses: L. Ge, Xerion Pharmeceuticals, Fraunhoferstr. 9, 82152 Martinsried, Germany; P. Pack, MTM Laboratories AG, Heidelberg, Germany.
E-mail address of the corresponding author: knappik@morphosys.de

0022-2836/00/60057–30 $35.00/0

see Winter *et al.*, 1994; Hoogenboom *et al.*, 1998; Spada *et al.*, 1997; Rodi & Makowski, 1999). Phage display was developed first (Smith, 1985) and has been improved the furthest, especially in the antibody field. It is likely that conventional hybridoma technology may be superceded by a combination of these technologies, since these approaches are faster, involve no animals, yield antibodies of at least comparable affinities and work also with self-antigens or toxic molecules (Hoogenboom *et al.*, 1998). The selection of antibodies must start from an initial, highly diverse library. Here, we describe the construction of such a library by total gene synthesis, based on a structural analysis of the human antibody repertoire.

Human antibodies are of particular interest, since they are considered to be valuable for therapeutic applications (Carter & Merchant, 1997), avoiding the HAMA (human anti-mouse antibody) response frequently observed with rodent antibodies. Although it has been demonstrated in many examples (Dall'Acqua & Carter, 1998) that chimerization or humanization of rodent antibodies through protein engineering can successfully retain the affinity and specificity of the parental molecule (Baca *et al.*, 1997), this strategy is time-consuming and still does not yield fully human antibodies.

Previous phage-display libraries of human antibodies have been generated from immunized donors (Barbas & Burton, 1996), germline sequences (Griffiths *et al.*, 1994) or, most recently, naive B-cell Ig repertoires (Vaughan *et al.*, 1996; Sheets *et al.*, 1998; De Haard *et al.*, 1999). Selection from these libraries by phage-display has yielded human antibodies against numerous haptens, peptides and proteins. While these libraries have all been successful, their uncontrollable composition and problems with the subsequent expression of the antibodies (see below) and restricted engineering possibilities made it desirable to use a complete protein engineering approach to solve the problem.

The success of obtaining high-affinity antibodies is generally assumed to be related to the initial library size (Perelson, 1989), even though the exact relation may not be tractable by theoretical considerations, as it may be antigen-dependent. Consequently, successful "one-pot" libraries have all been large (Griffiths *et al.*, 1994; Vaughan *et al.*, 1996; Sheets *et al.*, 1998; De Haard *et al.*, 1999). It is important to note that, obviously, only the functional library size, i.e. the number of correctly assembled clones without any frameshift, stop codon or deletion, will contribute to the diversity. This number can be orders of magnitude below the apparent diversity usually reported, which is normally obtained by counting the numbers of transformants.

It has been shown that the *Escherichia coli* expression yields of functional antibody fragments can vary dramatically, even if the antibody gene is expressed in the same format, vector and expression strain. This effect has been shown to depend on cellular folding, which in turn is influenced by the antibody sequence and can be successfully improved by protein engineering (Knappik & Plückthun, 1995). There is growing evidence that critical amino acid residues located in turns at the surface or at the variable-constant (V-C) interface are responsible for the misfolding, aggregation or even toxic effects on the *E. coli* cells, hence leading to poor expression yields. Mutating those residues improved expression titers several-fold, without adversely affecting the binding properties (Deng *et al.*, 1994; Knappik & Plückthun, 1995; Ulrich *et al.*, 1995; Jung & Plückthun, 1997; Nieba *et al.*, 1997; Forsberg *et al.*, 1997). As phage display depends on correctly folded antibodies, there is some selection against poor folders (Deng *et al.*, 1994; Jackson *et al.*, 1995; Jung & Plückthun, 1997; Bothmann & Plückthun, 1998), and thus the functional library size will be decreased. However, the selection is clearly not stringent enough to secure that all molecules selected from a phage display library will have acceptable folding properties. Thus, to maintain diversity and secure reasonable expression properties of the selected molecules, it would be advantageous to create antibody libraries starting from well-expressed frameworks. While such approaches have been reported (Pini *et al.*, 1998; Jirholt *et al.*, 1998), only single frameworks have been used in these attempts, and consequently, the structural diversity does not approach that of other naive libraries.

The humoral immune system, however, does not work by the "single-pot" approach (Nissim *et al.*, 1994), but rather uses an evolutionary strategy. The initial, antigen-independent variability is first generated during B-cell development by gene rearrangements (V(D)J-joining), leading to more than 10^9 different molecules at any one time in a human being (Winter, 1998). After a B-cell is activated, the antigen-driven process of somatic mutation is initiated (Rajewsky, 1996), and remarkable improvements in binding can be found. It has been shown that mutations occurring in CDRs 1 and 2 are preferentially selected (Wagner & Neuberger, 1996; Ignatovich *et al.*, 1997; Green *et al.*, 1998), as their diversity in the initial germline variants is much more limited than that of the CDR3s (Tomlinson *et al.*, 1996). The design of an artificial library should make it convenient to follow this same approach. Indeed, previous experiments with peptides (Cwirla *et al.*, 1997), RNA-aptamers (He *et al.*, 1996) and antibodies (Schier *et al.*, 1996a; Hanes *et al.*, 1998) have shown that the evolutionary approach and, in the case of antibodies, CDR walking (Yang *et al.*, 1995; Schier *et al.*, 1996a; Wu *et al.*, 1998) can dramatically improve affinities. However, in the absence of suitably engineered genes, such an optimization can be extremely laborious.

The human antibody germline repertoire has recently been completely sequenced. There are about 50 functional V_H germline genes located on chromosome 14 (Tomlinson *et al.*, 1992; Matsuda

& Honjo, 1996), which can be grouped into six subfamilies according to sequence homology. About 40 functional V_L kappa genes comprising seven subfamilies are located on chromosome 2 (Cox *et al.*, 1994; Barbie & Lefranc, 1998), and about 30 functional V_L lambda genes grouped into ten subfamilies can be found on chromosome 22 (Williams *et al.*, 1996; Kawasaki *et al.*, 1997; Pallares *et al.*, 1998). The groups vary in size from one member (e.g. V_H6 and Vκ4) to up to 22 members (V_H3), and the members of each group share a high degree of sequence homology. By comparing rearranged sequences of human antibodies with their germline counterparts we (this work) and others (Cox *et al.*, 1994; Ignatovich *et al.*, 1997) have found that many human germline genes are never or only very rarely used during an immune response.

In structural terms, the V_H and V_L domains comprising the antigen binding Fv moiety (see Figure 1) share a common fold that, in its central portions, is almost perfectly superimposable, even when fragments from different species are compared (Chothia *et al.*, 1998). Larger differences are observed only in the conformation of the CDRs, and it has been shown in a series of studies (Chothia & Lesk, 1987; Chothia *et al.*, 1989; Al-Lazikani *et al.*, 1997) that all CDRs except V_H CDR3 adopt only a few distinct conformations. Hence the repertoire of conformations is limited to a relatively small number of discrete structural classes, depending on both the CDR length and the so-called canonical amino acid residues (Chothia & Lesk, 1987).

Here, we report the design, construction and analysis of a novel human antibody library concept designated HuCAL (Human Combinatorial Antibody Libraries). Each of the human V_H and V_L subfamilies that is frequently used during an immune response is represented by one consensus framework, resulting in seven HuCAL master genes for heavy chains and seven for light chains, and thus 49 combinations. All genes were made by total synthesis, thereby taking into consideration codon usage, unfavorable residues that promote protein aggregation as well as unique and general restriction sites flanking all CDRs, leading to modular genes that contain readily accessible CDRs and can be easily converted into different antibody formats.

A first set of antibody libraries based on the HuCAL concept was created by randomizing both the V_H and V_L CDR3 encoding regions of the 49 master genes using trinucleotide cassette mutagenesis (Virnekäs *et al.*, 1994), which leads to high-quality libraries. The cassettes were designed such that the naturally occuring diversity was covered, both in terms of length and amino acid composition. The final HuCAL antibody libraries (HuCAL version 1) were extensively characterized by sequencing, expression behavior and numerous selection experiments against a wide variety of antigens.

Results

Analysis of the human antibody repertoire

Sequence analysis

Amino acid sequences from variable domains of human immunoglobulins were collected from Kabat (Kabat *et al.*, 1991; Johnson *et al.*, 1996;†) and Genbank (Benson *et al.*, 1997) and incorporated into three databases, V heavy chain (V_H), V kappa (Vκ) and V lambda (Vλ), and aligned, using the Kabat numbering system. For each of the three chain types, rearranged sequences were collected whenever more than 70 positions had been determined, giving 386, 149 and 675 entries for Vκ, Vλ and V_H, respectively, at the time of library design. Similarly, all germline sequences were collected (48, 26 and 43 entries for Vκ, Vλ and V_H, respectively), as the complete loci‡ (see Cook & Tomlinson, 1995), had not been published at that time. Finally, all known D and J sequences were collected. Although the design was started before the complete germline repertoire was known, the availability of the whole repertoire and a larger number of rearranged sequences would not have influenced the library design, which was demonstrated by repeating the analysis using the complete germline repertoire and a larger database (846, 413 and 1201 entries for Vκ, Vλ and V_H, respectively) of human rearranged sequences (see Figure 2).

The binning into families is somewhat arbitrary, depending on how the homology cutoff between families is defined. Initially, for Vκ, seven families were established, Vλ was divided into eight families and V_H into six families. The single V_H germline gene of the V_H7 family (van Dijk *et al.*, 1993) was included in the V_H1 family, since the genes of the two families are highly homologous. Upon more detailed analysis, regarding canonical CDR conformations and canonical framework residues as well as gene usage (see below), the number of families was raised to seven for V_H, but was reduced to four for Vκ and three for Vλ.

To further examine the concept of constructing HuCAL using the equidistant partitioning of sequence space as an efficient means to engineer library diversity, it was important to test the usage of the structural groups in actual rearranged genes of antibodies. By counting the number of differences between each rearranged entry and each germline sequence, the nearest germline counterpart was identified for each rearranged sequence. Altogether, 532 (79%) V_H sequences and 474 (86%) V_L sequences (343 Vκ and 131 Vλ) could be clearly assigned to germline counterparts.

† http://immuno.bme.nwu.edu/

‡ Now available at VBASE, http://www.mrc-cpe.cam.ac.uk/imt-doc/public/INTRO.html

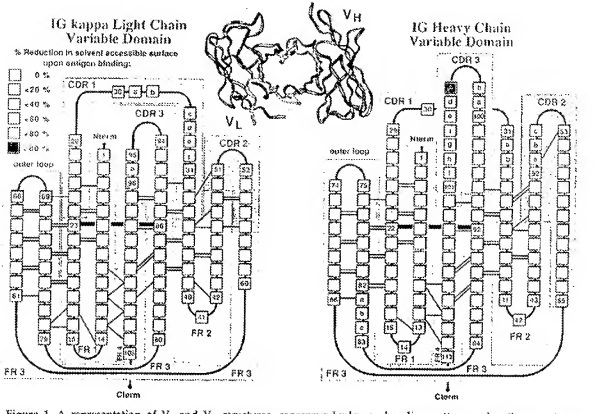

Figure 1. A representation of V_L and V_H structures, consensus hydrogen bonding pattern and antigen contacts. Residues are color-coded (from white to red) to indicate the reduction of residue solvent-accessible surface upon antigen binding, averaged over 52 liganded structures taken from the Brookhaven protein structure database (http://www.rcsb.org/pdb/). Residue numbering is according to Chothia *et al.* (1992), Tomlinson *et al.* (1995) and Williams *et al.* (1996), and CDR definitions conform to Kabat *et al.* (1991). Complementarity determining regions CDR1, CDR2 and CDR3 are indicated by blue, green and pink coloring, and framework regions by gray underlays.

Our results (see Table 1 and Figure 2) confirm the biased usage of human germline genes analysed previously (Tomlinson *et al.*, 1992; Cox *et al.*, 1994; Ignatovich *et al.*, 1997). The V_H germline gene usage was found to be restricted to about 12 genes from five sub-families, which are used in approximately 80% of all cases. The V_H2 family is only rarely used. Only four of the Vκ germline families were found to be used, and out of these only seven genes were used frequently (81%). The Vλ germline gene usage was found to be restricted to three families, which are used in 93% of all cases, and five genes from these three families were used most frequently (Table 1). We concluded that the vast majority (98% of all V_H, more than 99% of all Vκ and more than 93% of all Vλ) of human antibodies are derived from only five V_H and seven V_L families (four Vκ and three Vλ). Although the three germline genes of the V_H2 family are not frequently used, we decided to cover all six V_H families with our consensus approach, and therefore we included this family for further analysis.

The strategy of the synthetic library approach was therefore to represent each family by one representative member, subject to verification of the structural consequence of the distribution of CDR conformations (see the next section).

Structural analysis

Despite their great variability in length and sequence, the conformation of the antigen binding loops, denoted CDR (complementarity determining regions), have been shown to adopt only a limited number of main-chain conformations, termed canonical structures (Chothia *et al.*, 1989). The adopted structure depends on both the CDR length and the identity of certain key amino acid residues, both in the CDR and in the contacting framework, involved in its packing. The six V_H, four Vκ and three Vλ germline families, as defined above from the dendrogram analysis, were therefore analyzed for the canonical structures of CDRs that they were predicted to encode, in order to define the structural repertoire covered by these families (Table 1). In

Table 1. Frequency of germline family usage and corresponding types of canonical structures

Subfamily	Family usage (%)	Frequently used germline genes			Canonical structure prediction		Chosen HuCAL canonical structures	
		Locus	DP name	Usage (%)	CDR1	CDR2	CDR1	CDR2
VH1	19	1-69	DP-10	6		H2-2		H2-2
		1-18	DP-14	4	H1-1	H2-3	H1-1	H2-3
		1-02	DP-8	4				
VH2	2	-	-	-	H1-3	H2-1	H1-3	H2-1
		3-23	DP-47	12		H2-1		
VH3	34	3-30.3	DP-46	5	H1-1	H2-3	H1-1	H2-3
		3-48	DP-51	3		H2-4		
		4-34	DP-63	5	H1-1			
VH4	12				H1-2	H2-1	H1-1	H2-1
		4-59	DP-71	4	H1-3			
		5-51	DP-73	16				
VH5	19				H1-1	H2-2	H1-1	H2-2
		5-a	-	3				
VH6	14	6-01	DP-74	14	H1-3	H2-5	H1-3	H2-5
		O12	DPK9	9	L1-2	L2-1	L1-2	L2-1
Vκ1	32							
		O8	DPK1	7				
Vκ2	7	A3	DPK15	4	L1-3	L2-1	L1-4	L2-1
					L1-4			
		A27	DPK22	29	L1-2			
Vκ3	31	L6		11	L1-6	L2-1	L1-6	L2-1
		L2	DPK21	10				
Vκ4	10	B3	DPK24	10	L1-3	L2-1	L1-3	L2-1
Vκ5-7	0	-	-	-	L1-2	L2-1	-	-
Vλ1	31	1b	DPL5	13	13			
					14	7	13	7
		1c	DPL2	11				
		2a2	DPL11	18				
Vλ2	33				14	7	14	7
		2e	DPL12	11				
Vλ3	29	3r	DPL23	15	11	7	11	7
					12	7		
Vλ4-10	8	-	-	-	13	11	-	-
					14	12		

The human immunoglobulin germline subfamilies are listed together with their percentage usage as calculated by comparison with rearranged sequences. The percentage usage is determined from using the initial database of rearranged sequences with 1606 entries. The percentage usage calculated from the updated database with 2460 entries is given in Figure 2. The most frequently used germline genes according to our analysis are also given (locus name as well as DP nomenclature, see Tomlinson *et al.* (1992) for V_H, Cox *et al.* (1994) for Vκ, and Williams *et al.* (1996) for Vλ) together with their corresponding usage (derived from analysis of the smaller database). For details of the calculation, see the text. The canonical conformations that are present in each subfamily are shown together with the canonical conformations that have been chosen for HuCAL design. The canonical structure nomenclature is according to Chothia *et al.* (1992) for V_H, Tomlinson *et al.* (1995) for Vκ, and Williams *et al.* (1996) for Vλ.

the following, we will use the CDR definitions given by Kabat *et al.* (1991) (see also Figure 1) and the sequence numbering according to structural criteria defined by Chothia (Chothia *et al.*, 1992; Tomlinson *et al.*, 1995; Williams *et al.*, 1996).

The structural repertoire of the human V_H sequences was previously analyzed in detail by Chothia *et al.* (1992). In total, three conformations of CDR1 (H1-1, H1-2 and H1-3) and five conformations of CDR2 (H2-1, H2-2, H2-3, H2-4 and H2-5) have been defined, and the observed combinations have led to the conclusion that almost all sequences have one of seven main-chain folds. For the highly diverse CDR3, which is encoded by the D and J-minigene segments and uncoded nucleotides (N-region diversity), structural families have been defined only very recently (Morea *et al.*, 1998; Oliva *et al.*, 1998), but structural predictions are not approaching the accuracy seen for the canonical folds of the other CDRs.

All members of the V_H1 family encode the CDR1 conformation H1-1, but differ in their CDR2 conformation: both the H2-2 and the H2-3 conformation were found in five germline genes. Since these two types of CDR2 conformations are defined by different types of amino acids at position 71 located in framework 3, we divided the V_H1 sub-family into two further sub-families: V_H1A with CDR2 conformation H2-2 (alanine at position 71) and V_H1B with the conformation H2-3 (arginine at position 71). Upon model building (see below), we decided to include both gene types into the library design and to construct both a V_H1A and V_H1B master gene (see below).

The members of the V_H2 family were all predicted to have the conformations H1-3 and H2-1 in CDR1 and CDR2, respectively.

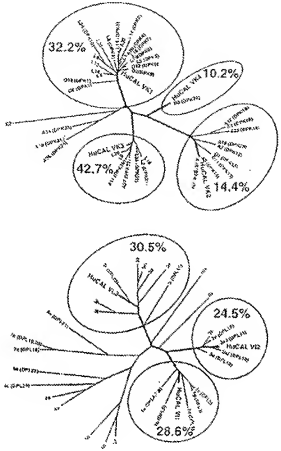

Figure 2. Coverage of germline sequence space by HuCAL sequences. The protein sequences representing the human V_L and V_H germlines were taken from VBase (http://www.mrc-cpe.cam.ac.uk/imt-doc/public/INTRO.html) and aligned to the 14 HuCAL sequences. The Phylip (http://evolution.genetics.washington.edu/phylip.html) and ClustalW (see ftp://ftp.ebi.ac.uk/pub/software/mac/clustalw/) phylogeny program packages were used to generate separate unrooted trees for the V_L kappa, V_L lambda and V_H sequences. Percentages indicate the fraction of rearranged sequences in the database that cluster within the different germline subgroups. For these calculations, we used a database of rearranged sequences with 846 V_L kappa, 413 V_L lambda and 1201 V_H sequence entries. The difference to 100% in the case of V_L kappa (0.5%) and lambda (16.4%) is due to rarely used germline subfamilies that are not represented by the HuCAL master genes.

The CDR1 conformation of the V_H3 family members was predicted in all cases to be H1-1, but three different types were found for CDR2 (H2-1, H2-3, H2-4). In these CDR2 conformations, the canonical framework residue 71 is always arginine, while the loop conformation of CDR2 is defined by

the residues 52a and 55 as well as the length variation. Of the rearranged V_H3 sequences, 80% were predicted to contain the H2-3 conformation. Therefore, the V_H3 family is best represented by a sequence containing the canonical conformations H1-1 and H2-3, even though the more groove-like shapes of binding sites from the longer CDR-H2 types may be introduced later by CDR shuffling.

The V_H4 family members were predicted to contain three types of CDR1 conformations; namely, H1-1, H1-2 and H1-3. The CDR1 canonical framework residue 26 was found to be glycine in all cases, and the CDR1 loop conformation is defined solely by residues located in this region. Since 62% of all rearranged V_H4 sequences contained the H1-1 type of CDR1, this conformation was chosen for representing the V_H4 family. The CDR2 conformation of the V_H4 members was found to be H2-1 in all cases.

The two members of the V_H5 family were found to have the conformation H1-1 and H2-2, and the single germline gene of the V_H6 family had the conformation H1-3 and H2-5 in CDR1 and CDR2, respectively. Hence, in structural terms the majority of the frequently used members of the six V_H families can be represented by seven sequences, since only the V_H1 family contained two types of canonical CDR folds defined by residues in the framework region, and since V_H3 and V_H4 were decided to be represented by the most prevalent type. The canonical conformations not present in the design can be incorporated later during CDR library generation, since the key residues for those conformations are part of the CDR itself.

The structural repertoire of the human Vκ germline sequences was analyzed by Tomlinson *et al.* (1995). There are four conformations of the CDR1, which are defined by the length of the loop (7, 8, 12 and 13 amino acid residues) and the nature of residues 2, 25, 29, 33 and 71. The CDR2 loop of human Vκ domains is only three amino acid residues long in all cases, and is predicted to adopt a single canonical fold. Most human Vκ germline segments encode also a single conformation of the CDR3 loop, which is stabilized by the conserved *cis*-proline 95, but other conformations in rearranged sequences are possible due to the process of V-J joining and the potential loss of this proline residue. Since the CDR3 region was planned to be randomized for library generation, this area was not considered for the consensus sequence design. Hence, the structural repertoire of Vκ domains is essentially defined by the conformation of the CDR1 region. All members of the Vκ1 family contained a seven residue CDR1 (L1-2), and the most frequently used members of the Vκ2 family contained a 12 residue CDR1 (L1-4). The members of the Vκ3 family contained either a seven (L1-2) or an eight (L1-6) residue CDR1. Since the canonical framework residues that additionally define the CDR1 conformation are identical in both cases, and since more than 60% of the rearranged Vκ3 sequences contained the CDR1 conformation L1-6, this type was chosen for the consensus sequence. The single germline member of the Vκ4 family contained a 13 residue CDR1 (L1-3).

The structural repertoire of the human Vλ germline sequences was analyzed by Williams *et al.* (1996). The three families analyzed here encode identical conformations of the CDR2 loop. The CDR3 loop conformation is thought to be more highly variable, as there is some length variation and no *cis*-proline residue. Since this part was planned to be randomized for library generation, this area was not considered for the consensus sequence design. Although the CDR1 region of the Vλ1 family contains either 13 or 14 amino acid residues, it is thought to adopt a single conformation, since the canonical key residues are conserved and the additional insertion of one residue has little effect on the overall structure (Chothia & Lesk, 1987). A CDR1 length of 13 residues, which was found in more than 90% of all rearranged Vλ1 sequences, was chosen for the Vλ1 consensus. The members of the Vλ2 and Vλ3 families each encode a single defined type of CDR1 loop structure: the Vλ2 family encode a CDR1 loop of 14 residues, and the CDR1 loop length of the Vλ3 family is 11 residues.

In summary, from the eight different pairs of CDR1-CDR2 conformations encoded by the Vκ and Vλ germline genes that are used frequently, seven could be represented by four Vκ and three Vλ consensus genes. The remaining CDR1 conformation (seven residue CDR1 loop in the Vκ3 family) is not defined by canonical key residues in the framework region and can therefore be inserted into the Vκ3 consensus sequence during library generation. From the 11 different family-specific pairs of CDR1-CDR2 conformations found in the six V_H germline families, seven could be covered by dividing the family V_H1 into two families (V_H1A and V_H1B). The remaining four pairs (two in the V_H3 and two in the V_H4 family) were either not found in rearranged antibody sequences or are defined by the CDRs themselves and will therefore have to be created during the construction of CDR libraries. Hence, the structural repertoire of the human V genes used could be covered by 49 ($7\ V_H \times 7\ V_L$) different frameworks.

Design of consensus frameworks

The compilation of rearranged sequences was first divided into separate groups (four Vκ, three Vλ and seven V_H) according to the germline families described above. These protein sequence databases were used to compute the consensus sequences of each subgroup. By using the rearranged sequences instead of the germline sequences for calculating the consensus, the consensus was automatically weighted according to the frequency of usage. Additionally, frequently mutated and highly conserved positions could be identified.

For the CDR1 and CDR2 regions, the consensus of rearranged sequences was replaced with the amino acid sequence of one of the germline sequences of the corresponding family. This procedure removes any bias, as the CDRs of rearranged and mutated sequences are known to be mutated due to selection towards their particular antigens. In the case of Vλ, a few amino acid exchanges were introduced in some of the chosen germline CDRs in order to avoid structural constraints (position 30b in Vλ1 and positions 27 and 34 in Vλ3, see Figure 3).

To construct, assemble and verify the genes, as well as to obtain preliminary information on expression behavior, it was advantageous to first substitute the intended library of CDR3-H and CDR3-L cassettes with defined dummy sequences. We chose the sequences $_{95}$QQHYTTPP and $_{95}$WGGDGFYAMDY for the V_L and V_H chains, respectively, which are derived from the antibody 4D5 (Carter *et al.*, 1992a) and are known to be favorable for antibody folding in *E. coli* (Jung & Plückthun, 1997). Even though molecular modeling indicates that the omega loop from Vκ is not ideal in a Vλ framework because of steric clashes, good expression behavior could still be obtained, demonstrating the robustness of the frameworks (see below)

For the framework 4 regions, encoded by the J-elements, the consensus of the rearranged sequences in each family was calculated and found to be identical in all families of V_H and V_L (κ and λ). This shows that there is no correlation between V-usage and J-usage (Baskin *et al.*, 1998). In all three cases, this consensus sequence was identical with at least one of the naturally occurring sequences encoded by joining elements, indicating that the sequence is able to exist.

We have described, up to this point, only sequence information that was used to design the consensus sequences. It could therefore not be excluded that the consensus would lead to a molecule whose sequence might "jump" between different naturally occurring sequences, thereby creating certain artificial combinations of amino acid residues that are located far away in the sequence but give rise to contacts in the three-dimensional structure. It was therefore essential to verify the sequences by structural means. Otherwise, the uncritical use of the algebraic consensus might obscure a hidden interaction between certain residues, which can occur only in certain combinations. While this approach may also keep residues together that are linked only historically, it does safeguard against losing hidden long-range interactions (Saul & Poljak, 1993). As a first check, the most homologous rearranged sequence for each consensus sequence was identified by searching against the compilation of rearranged sequences, and all positions where the consensus differed from this nearest rearranged sequence were inspected (see Materials and Methods). Furthermore, models for the seven V_H and seven V_L consensus sequences were built and analyzed according to their structural properties (see the next section). As a result of this analysis, the following residues were exchanged (given is the position according to Kabat's numbering scheme, the substitution performed, and the name of the gene family): $S_{76.5}T$ (V_H2), $N_{1.34}A$ (Vκ1), $G_{1.6}A$, $D_{1.60}A$, $R_{1.7}S$ (Vκ3) and $V_{1.7}T$ (Vλ3).

After the consensus protein sequences were designed, phylogenetic trees were built with the programs PHYLIP† and ClustalW‡ (Thompson *et al.*, 1994). For this representation, we repeated the analysis of germline usage based on an updated database of rearranged human antibody sequences that was more than twice the size of the original database that we used for the design of the HuCAL sequences. Separate unrooted trees were built for the V_Lκ, V_Lλ and V_H sequences (Figure 2). This analysis illustrates the strategy adopted in the present study, which is an attempt to approach a more equidistant representation of sequence space, by having only one member for each of the main "branches" of the tree. By analyzing each consensus sequence as if it were a member of the germline, its position in the sequence map is indicated, and that it truly represents the family (Figure 2).

Molecular modeling and analysis

To obtain more information about the packing, CDR conformations and framework properties, all seven V_H frameworks, all four Vκ frameworks and the three Vλ frameworks were built *in silico* homology modeling. As a basis, a complete structural alignment of the approximately 100 independent antibody sequences available in the PDB (Bernstein *et al.*, 1977) was carried out as indicated in the legend to Figure 3. Usually, the template with the highest resolution and the fewest mutations relative to the consensus sequence to be modeled was used. For all models, multiple templates were compared, such that the effect of mutations in any of the templates could be evaluated directly from the structural alignment. The experimental structures displaying the highest degree of similarity to each of the HuCAL constructs are listed in Table 1 of the Supplementary Material.

In the models (see Figure 4), the dummy CDR3 sequences from the antibody hu4D5 (version 8) are shown (PDB file 1FVC). All models were checked with the program PROCHECK§ (Morris *et al.*, 1992; Laskowski *et al.*, 1993) and were shown to have no more residues in the less favorable regions of the Ramachandran plot than the template structures. (some unfavorable torsion angles in loop regions

† see http://evolution.genetics.washington.edu/phylip.html

‡ ftp://ftp.ebi.ac.uk

§ http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html

are highly conserved, e.g. position 51 at the tip of CDR2 in V_L), as well as having no obvious cavity or unusual exposed hydrophobic region, and a full set of standard variable domain hydrogen bonds.

Consistent with sequence considerations, the great majority of canonical structures was predicted to be present by model building, when comparing the critical residues with the templates. More recent work (unpublished results), based on previous experimental observations from X-ray crystallography (Saul & Poljak, 1993) and mutagenesis (Langedijk *et al.*, 1998), has uncovered several more structural relationships within each V_H domain, which may to contribute to diversity. Particularly, relationships between the nature of the residues H6, H7 and H9, due to the different hydrogen bonding pattern of H6 to the backbone, can transmit a conformational change through the protein *via* residues H18, H82, H67, and H63. Our analysis showed that all types of conformations that occur commonly in natural human frameworks are represented in the chosen consensus frameworks.

In the V_H3 group of germline sequences, there is more variation in CDR2, because of the length variation of a two amino acid residue insertion occurring in a group of human sequences (positions 52b and c). These antibodies might form more cleft-like binding pockets, and this diversity is not present in the original library design, even though many other combinations of frameworks would be able to form cavities and clefts as well. Through the modular design, however, these longer CDR2 elements can easily be introduced by cassette mutagenesis.

An analysis in analogy to that reported by Nieba *et al.* (1997) showed that the exposed residues at the V/C interface are already of low hydrophobicity in all consensus frameworks, consistent with their superior expression behavior in *E. coli* (see below). Moreover, many of the residues identified as crucial for stability and clearly selectable by phage display, such as P_{L8} (defining a conserved kink in the first β-strand with a *cis*-peptide bond in Vκ domains, or the *trans*-proline residues at positions 8 and/or 9 in Vλ domains, see Spada *et al.*, 1998) are present in all master sequences. Residue R_{H66}, which is part of a conserved charge cluster, and frequently K in murine antibodies, where it leads to lower stability (see Proba *et al.*, 1998), is present in all master genes except V_H5, where the consensus was found to be Q_{H66}. All residues known to make conserved side-chain hydrogen bonds are present in the master genes. Side-chain to side-chain: K_{H28} to Q_{H66}, D_{H86} and Y_{H66}; R_{H66} to D_{H82}; R_{H94} to D_{H101}; Q_{L6} to T_{L107}; Q_{L37} (L_{L37} in Vκ2) to Y_{L87}; R_{L61} to D_{L82}. Side-chain to main-chain CO: R_{H66} to H_{82a}; T_{H87} to X_{H89}; Q_{L6} to X_{L86}; Q_{L38} to X_{L42}. Main-chain NH to side-chain: X_{H66} to Y_{H90}; X_{H77} to D_{H77}; X_{H83} to D_{H86}; X_{H89} and X_{H104} to E_{H6} or Q_{H6}; X_{H111} to T_{H87}; X_{L79} to D_{L82}; X_{L84} and X_{L101} to Q_{L6}. Interdomain: Q_{L38} to Q_{H39}. In this listing, X refers to positions without dominant residue preference.

The relative orientation of V_L with respect to V_H is still understood only poorly, and will depend on the exact pairwise combination and on the specific CDR3 sequences. Frequently, monoclonal antibodies are found with mutations within the interface. This introduces further uncertainty in building a model of the combining site, because a small deviation in angle can have a large effect at the top of the binding site. This variability of the relative orientation of the two domains is particularly large for Vλ domains and Vκ lacking the *cis*-Pro in position L95, and is further modulated by non-tyrosine residues in position L49. The "elbow" of ordinary Vκ CDR3 inserts around L96 into a notch in V_H and restricts the flexibility of the interface. Since the interface residues are highly conserved between all the consensus antibodies (see Figure 3), and since very similar frameworks are available as templates in the database, more reliable models may be possible for HuCAL antibodies than for antibodies further away from the consensus. This system of defined frameworks might, in addition, provide excellent access to studying this question of domain orientation experimentally.

Construction of the seven V_H and seven V_L master genes

The final result of the analysis described above was a collection of 14 amino acid sequences, which represent the frequently used antibody repertoire of the human immune system. These sequences were then back-translated into DNA sequences. In a first step, the back-translation was carried out using only codons that are known to be used frequently in *E. coli*. In a second step, these gene sequences were then examined for all possible restriction endonuclease sites, which could be introduced without changing the corresponding amino acid sequences. This was done by creating a database of all possible silent cleavage sites for each gene. Using this database, cleavage sites were selected that were located close to the CDR and framework borders and that could be introduced into all V_H, Vκ or Vλ genes simultaneously at the same position. This was considered essential to the overall strategy, as CDRs (or frameworks) can then be shuffled within pools of sequences, without even knowing the individual antibody sequence. In a few cases it was not possible to find a common cleavage site for all genes at one of the flanking regions. In that case, one amino acid residue of the sequence was changed if this change seemed to be feasible according to the available sequence and structural information as delineated in the molecular modeling section. Each sequence was then analyzed again after exchange as described above.

In total, six amino acid residues were exchanged during the design of the genes: $T_{H28}Q$ (V_H2), $S_{H07}G$ (V_H6), $E_{L1}D$ and $I_{L58}V$ (Vκ3), $K_{L22}R$ (Vκ4) and $T_{L72}S$ (Vλ3). Additionally, the first two amino acid residues of all three Vλ sequences were changed to

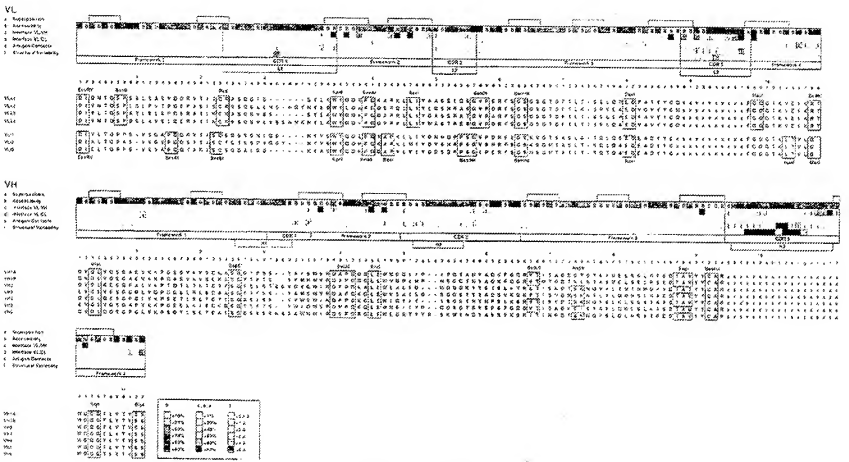

Figure 3. (*Legend opposite*)

aspartate-isoleucine in order to introduce an *Eco*RV site common to all V_L genes. After this design, only one element junction remained where no common cleavage site could be found. For this region (the border between CDR2 and framework 3 in the V_H sequences), two different types of cleavage sites were used instead: *Bst*EII for V_H1A, V_H1B, V_H4 and V_H5; and NspV for V_H2, V_H3, V_H4 and V_H6.

During this analysis, several potential restriction endonuclease sites were identified that could be introduced into every one of a given group without changing the amino acid sequence, but which were not located at the flanking regions of the CDR or framework elements. The introduction of these cleavage sites made the system more flexible for further improvements. Finally, each gene sequence was modified again to remove, with the exception of the common restriction sites, all but one of the other sites (with a length of the recognition site of five or more bases), since this unique site might be used as a "fingerprint site" to differentiate the genes by restriction digest. All these changes were again carried out without changing the corresponding amino acid sequence. The 14 final protein sequences, including the introduced restriction pattern are shown in Figure 3.

The resulting consensus protein sequences were finally compared to the germline sequences, and a mean deviation of all 49 consensus sequences from their closest germline counterparts of 4.9(±3.6) residues was found. Thus, these consensus sequences are, on average, much more related to the germline sequences than the majority of rearranged sequences found in the database (mean deviation 14.7 amino acid residues). In contrast to the "original" germline sequences, however, our synthetic versions have all the advantages of sequences with known and predictable unique restriction sites at the framework/CDR borders.

The consensus gene fragments were then assembled from oligonucleotides by SOE-PCR assembly (see Materials and Methods for details). Gene segments encoding the human constant domains C_H1 (sub type IgG1), Cκ and Cλ1 were designed with optimized *E. coli* usage and synthesized in order to create F_{ab} fragments for display or expression (see Materials and Methods). After synthesis, the gene fragments were assembled and inserted individually into the expression vector pBS12, yielding 49 single-chain Fv genes containing identical dummy V_H and V_L CDR3s. The general format of the scFv genes is shown in Figure 5. All 49 master genes were also cloned in the reverse oriented scFv format (V_L-V_H) as well as in the F_{ab} format for future libraries (data not shown).

E. coli expression analysis

The *E. coli* expression of the 49 scFv genes (all containing the same V_H and V_L CDR3s from the antibody hu4D5, see Carter *et al.*, 1992a) was studied similarly as described by Knappik & Plückthun (1995). We found that all 49 master genes could be expressed as soluble proteins in the periplasm of *E. coli*, yielding a band of the correct size in FLAG Western blots of soluble *E. coli* crude extracts (data not shown). This indicates that all 49 combinations are most likely capable of forming V_H/V_L pairs, since unpaired domains tend to aggregate (Wall & Plückthun, 1999).

The ratio of soluble to insoluble expressed protein was quantified from Western blot experiments for each scFv gene, since this value has been shown to be correlated to the expression behavior

Figure 3. Protein sequences of the HuCAL V_H and V_L master genes. An alignment of the seven V_L and seven V_H sequences is shown, together with the approximate location of restriction endonuclease sites that were introduced into the corresponding DNA sequences. The alignment, numbering and loop regions (L1-L3, H1-H3) are according to structural criteria defined by Chothia *et al.* (1992), Tomlinson *et al.* (1995) and Williams *et al.* (1996). The H3 loop is given as defined by Chothia & Lesk (1987), although more recently, the extended H3 loop has been defined to include residues 92 and 104 (Morea *et al.*, 1998). CDRs are according to Kabat *et al.* (1991). Color codes indicate: (a) the structurally least variable regions used for least-squares superposition of the C coordinates of structures and models (residues L3-L7, L20-L24, L33-L39, L43-L49, L62-L66, L71-L75, L84-L90 and L97-L103 for V_L; H3-H7, H19-H23, H34 to H40, H44-H50, H67-H71, H78-H82, H88-H94 and H102-H108 for V_H) indicated as gray bars. (b) The average relative side-chain solvent-accessibility in the isolated domains, indicating the average side-chain solvent-accessibility for each position: 100% indicates a solvent-accessible surface of the same side-chain in the context of a poly(Ala) peptide in extended conformation. Strongly buried positions (less than 30% of the side-chain surface is solvent-accessible) are additionally marked by B, semi-buried positions (less than 50% of the side-chain surface is solvent accessible) are additionally marked by b. (c) The average loss of side-chain solvent-accessible surface upon formation of the V_L/V_H dimer interface, indicating residues directly contributing to the dimer interface. Positions strongly buried upon interface formation (more than 80% of the residual solvent-accessible surface buried in the interface) are additionally marked by I, and semi-buried positions (more than 40% of the residual solvent-accessible surface buried in the interface) are additionally marked by i. (d) The average loss of side-chain solvent-accessible surface upon formation of the VL/CL and VH/CH interface in the Fab fragment. (e) Average loss of side-chain solvent-accessible surface upon binding of the antigen. Positions strongly buried upon antigen binding (more than 80% of the residual solvent accessible surface buried in the interface) are additionally marked by I, and semi-buried positions (more than 40% of the residual solvent accessible surface buried in the interface) are additionally marked by i. (f) Average deviation of the C$^\alpha$ positions of all V_L or V_H structures, respectively, in the PDB database (http://www.rcsb.org/pdb/) from the average C$^\alpha$ positions.

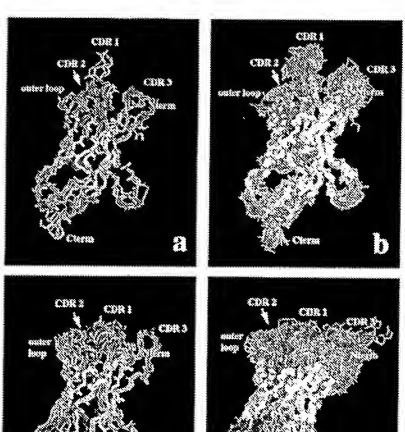

Figure 4. Coverage of the range of conformational variability of natural antibodies by the HuCAL frameworks. The homology models of the 14 HuCAL framework structures were generated using the program Insight II, modules Homology, Biopolymer and Discover (Biosym/MSI, San Diego, CA) as described in Materials and Methods. For CDR3, the sequence of antibody huID5 was used in all the models. The resulting V_L and V_H models were aligned by least-squares superposition of the C^α coordinates of residues L3-L7, L20-L24, L33-L39, L43-L49, L62-L66, L71-L75, L84-L90 and L97-L103 for V_L and H3-H7, H19-H23, H34 to H40, H44-H50, H67-H71, H78-H82, H88-H94 and H102-H108 for V_H (indicated in white). For comparison, 100 non-redundant V_L and V_H structures (mouse and human) were taken from the the RCSB protein structure database (http://www.rcsb.org/pdb/) and aligned. (a) HuCAL V_L models and (b) X-ray structures: cyan, kappa chains; blue, kappa chains lacking cis-Pro L8 (mouse only); pink, lambda chains. (c) HuCAL V_H models and (d) X-ray structures color-coded according to the sequence pattern correlating with the framework structure subtypes: magenta, H6 = Glu, H9 = Pro; pink, H6 = Glu, H9 = Gly; cyan, H6 = Gln, H9 = Ala; blue, H6 = Gln, H9 = Pro. The fourth conformation not covered by the HuCAL models shows some correlation with the presence of Pro in position H7, which is very rare in human sequences (<1 %), but frequently seen in mouse sequences (in about 22 % of the sequences).

of antibody fragments (Knappik & Plückthun, 1995; Nieba *et al.*, 1997; Jung & Plückthun, 1997). In each separate expression experiment, the HuCAL H3κ2 master gene was included as an internal control. The results are given in Table 2. The HuCAL genes were found to show a higher ratio of soluble to insoluble protein than many antibody genes obtained from natural monoclonal antibodies and subsequently expressed in *E. coli*. The ratio of soluble to insoluble protein ranges from 33 % ($H_1\lambda2$) to 90 % ($H_1\lambda\lambda1$ and $H_3\kappa1$), whereas a wide range of ratios has been found from natural antibody fragments, including many with ratios much below 30 % under similar experimental conditions (Forsberg *et al.*, 1997; Nieba *et al.*, 1997). We could not find a correlation

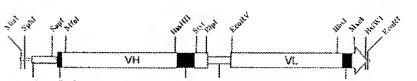

Figure 5. Arrangement of HuCAL scFv in the V_H-V_L orientation. The scFv gene cassette is preceded by a *phoA* signal sequence and a short FLAG tag. The two domains are fused by a 20 amino acid residue flexible linker. Some of the unique restriction sites common to all master genes are shown, and the location of the CDR3 regions is indicated.

between the type of V_L gene and expression behavior of the corresponding scFv genes, but it seemed that the genes encoding the V_H3 or V_H1A domains are showing higher soluble to insoluble ratios in almost all combinations (Table 2). These initial findings clearly need to be extended by a more detailed biophysical characterization.

The amounts of soluble protein produced, when compared to the H3κ2 gene set as 100%, ranged from 26% to 212% (data not shown), indicating that soluble expression yields for all combinations fall into a narrow range. Although we must expect that differences in the CDRs after randomization and selection of binders may influence the range of expression yields seen with the master genes, the use of well-expressed frameworks for creating libraries increases the chance to select well-expressed binding antibodies and reduces the large imbalances in the display efficiencies.

The CDR3 sequence introduced as dummy sequence in all V_L genes was taken from a V_L kappa gene (see above). Since this Vκ CDR3 contained a *cis*-proline residue at position 95, creating an omega-loop that is normally not found in V_L lambda CDR3s, and which might influence the folding and hence the expression behavior of the corresponding scFv genes, a Vλ dummy consensus CDR3 cassette encoding the sequence $_{89}$QSYDSSLS was designed and used to replace the Vκ dummy CDR3 in the H3λ1 scFv gene. Interestingly, however, no significant difference in expression yields could be detected (data not shown).

The expression behavior of two randomly chosen scFv genes (H2κ2 and H3κ2) was analyzed in more detail. These two genes were selected from panning experiments after library creation (see below) and therefore contained CDR3 sequences different from the dummy sequence of the master genes. Since both scFv fragments bound the antigen they were selected on, we could use ELISA experiments to determine the amount of active material in the lysates after different times of induction. The results are shown in Figure 6. We found that the expression titer after five hours of induction at 30 °C was 6 mg (H2κ2) and 10 mg (H3κ2) per liter of shaking-flask culture. The expression titer stayed constant for several hours and then decreased, probably due to the start of cell leakiness. This observed expression yield is significantly higher than that reported for antibody fragments from other libraries (Griffiths *et al.*, 1994; Vaughan *et al.*, 1996).

Design and construction of CDR3 library cassettes

Our rational approach to creating an antibody library aims at defining, with the smallest number of molecules possible, a structural diversity as large as possible. At the same time, it was important to design molecules that are likely to be stable and fold well. Furthermore, it was essential to direct the sequence diversity to those residues most likely in contact with the antigen. We decided for the first set of HuCAL libraries to randomize both CDR3 regions of the V_H and V_L genes simultaneously, since these two regions form the inner circle of the antigen binding site, and therefore show the highest frequency of antigen contacts in structurally known antibody-antigen complexes. In order to obtain the highest degree of diversity in

Table 2. Expression analysis of HuCAL master genes

	κ1	κ2	κ3	κ4	λ1	λ2	λ3
H1A	61	58	52	42	90	61	60
H1B	39	48	66	48	47	39	36
H2	47	57	46	49	37	36	45
H3	85	66 ± 6	76	61	80	71	83
H4	69	52	51	44	45	33	42
H5	49	49	46	67	54	46	47
H6	90	58	54	47	45	30	51

The amount of soluble full-length scFv relative to the total amount obtained is given (in %), as determined from quantitative Western blot analysis of HuCAL master genes expressed in *E. coli* after two hours of induction at 30 °C. The H3κ2 master gene which served as an internal control in each separate expression experiment was analyzed altogether 18 times. For this gene, the mean value and the standard deviation is given.

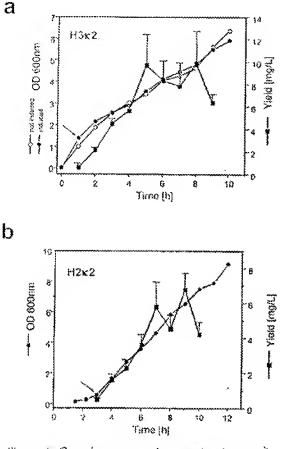

Figure 6. Growth curves and expression kinetics for two HuCAL scFv fragments. (a) Gene derived from the H3κ2 HuCAL framework; (b) gene derived from the H2κ2 HuCAL framework. The growth curves (circles) were determined by measuring the absorbance at 600 nm at the indicated time-points. For comparison, the growth curve of the uninduced culture (open circles) is given in (a). The arrows indicate the time-point of induction. The amount of functional scFv (squares) at the different time-points was determined by ELISA measurements of crude extracts. The corresponding purified antibody fragments of known concentration, also measured in the presence of a corresponding amount of cell extract, served as internal standard to calculate the scFv amount based on the ELISA signal obtained. The mean and standard deviation of three different measurements is given for each experiment.

the V_H CDR3, which is also the most variable region in natural antibodies, we applied the following strategy for library generation: first, we designed V_L CDR3 library cassettes strongly biased for the known natural distribution of amino acids (see below) with relatively low complexity and inserted those in the V_L master genes, aiming at a library size of about 10^7 members. Subsequently, we used these V_L libraries to insert a V_H CDR3 library cassette with very high complexity (both in terms of sequence composition and length variation), ensuring that every single library member contains a unique V_H CDR3 sequence.

Since we used trinucleotides (Virnekäs *et al.*, 1994) for the generation of the CDR3 library cassettes (see below), we could introduce any amino acid bias at any position of the cassettes. We decided to first analyze the sequence variability in the CDR3 regions of our databases of human rearranged antibody sequences and use this information together with structural data for the library design, in order to bias the CDR3 sequences towards the naturally found human antibodies.

Vκ CDR3

A total of 382 sequences of rearranged antibodies from our initial internal database were analyzed. In the following discussion, we will use the numbering system and definitions of CDRs regions introduced by Kabat even though this does not always correspond to the structural definitions (Chothia & Lesk, 1987; Barre *et al.*, 1994; Giudicelli *et al.*, 1997).

A fraction of 72.3 % of all CDR3s had a CDR length of eight amino acid residues, the remaining sequences had CDR lengths of less than seven (1.8 %), seven (7.3 %), nine (17.3 %), and ten (1.3 %) residues. Because of the predominance of CDRs of eight residues, we decided to consider just that size for constructing a CDR3 library. The omega-loop structure of Vκ CDR3 is determined by a characteristic *cis*-proline residue at position 95, which is encoded in 96 % of all κ germline genes, but can be lost upon V-J rearrangement. A total of six canonical structures have been discussed with structural data being available for structures 1 and 2 (Al-Lazikani *et al.*, 1997). In canonical structure 1, residues 90 and 95 are predominantly occupied by glutamine and proline, respectively, whereas in structure 2, the presence of *cis*-proline at position 94 is characteristic. About 87 % of all 382 sequences had Q_{L90}, and 78 % had P_{L95}, whereas P_{L94} was present in only 1 % of all sequences. Therefore, we decided to base the design of Vκ CDR3 on structure 1. Besides the canonical residues, position 89 showed a strong conservation, with glutamine present in 89 % of all sequences. Residues 89 and 90 are not part of the region outside the β-strand forming the CDR-L3 loop, which comprises residues 91 to 96 (Chothia & Lesk, 1987). Within CDR-L3, a high degree of variability (except for position 95 mentioned before) can be seen, with some preference for tyrosine at position 91. This corresponds well with the inspection of antigen contact residues in structurally known antibody-antigen complexes, showing that positions 91 to 94 and 96 seem to play the most important role (see Figure 3).

In our design of the library (see Figure 7(b)), we kept position Q_{L90} constant. Besides being a canonical residue, the side-chain of this glutamine residue does not contribute to the antigen-binding pocket, but points in the opposite direction. In the trinucleotide mixture, we biased positions 89 and

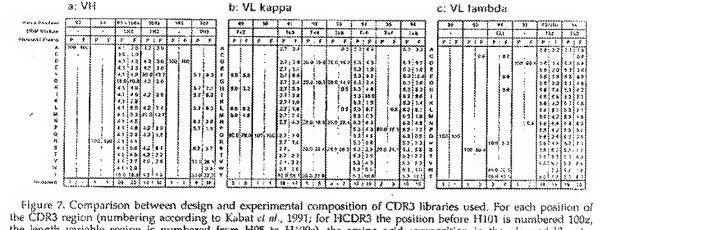

Figure 7. Comparison between design and experimental composition of CDR3 libraries used. For each position of the CDR3 region (numbering according to Kabat *et al.*, 1991; for HCDR3 the position before H101 is numbered 100z, the length variable region is numbered from H95 to H100s), the amino acid composition in the planned libraries (P, left columns) is compared with the composition found from sequencing 257 clones of the initial libraries (F, right columns). The TRIM mixture indicates the mixtures of trinucleotides used in the oligonucleotide synthesis (see Table 3 of the Supplementary Material). Occupied indicates the number of amino acids encoded by the respective mixture and found in the sequenced clones, respectively.

95 strongly towards glutamine and proline, respectively. A limited set of trinucleotide codons was allowed for positions 92 and 93, despite the fact that a large number of different residues can been found there, because the side-chains of these residues point away from the V_H CDR3 contact side. In contrast, for position 91, 18 amino acids (all except cysteine and proline) were allowed (biased towards Y_{L91}). Since proline is never found at position 91 in germline or rearranged sequences, it could be that P_{L91} would not allow the loop to form the correct conformation. Cysteine was omitted, since it was almost never found and it might cause problems during phage panning and later expression because of disulfide formation. Accordingly, for positions 94 and 96, all amino acids except cysteine were allowed. The residues located at those three most strongly randomized positions point into the binding pocket. By focusing the diversity towards positions that are most likely in contact with the antigen, we could reduce the overall theoretical diversity to a value of 1.3×10^9, which ensured that the theoretical diversity will be present in the final library.

For the four different HuCAL Vκ consensus genes, three trinucleotide-containing oligonucleotides were synthesized. A single oligonucleotide for Vκ1 and Vκ3 was possible, since both differ only at position 85 (κ1 T_{L85}; κ3 V_{L85}) and could thus be synthesized by using a mixture of two trinucleotides encoding threonine and valine in a 1:1 ratio at the appropriate position. Structural inspection revealed that residue 85 has no contact to other residues, thus making it likely that an exchange of these two similar amino acids would not cause problems. Indeed, we found the expected 1:1 ratio at this position after library construction and sequencing of clones (data not shown).

For oligonucleotide synthesis, six different trinucleotide mixtures (Tκ2 to Tκ6, see Figure 7(b)) had to be prepared comprising two to 19 codons, either biased or equally distributed. While initial results had suggested that different trinucleotides couple with different relative coupling yields (Virnekäs *et al.*, 1994), more controlled subsequent experimentation showed that these differences were not systematic (data not shown) and thus, trinucleotide mixtures were prepared directly using the desired molar ratios, thereby implicitely assuming an equal coupling yield. During oligonucleotide synthesis, the stepwise coupling ratio for trinucleotide mixtures ranged from 95.5 % to 97.5 %, the overall yield per oligonucleotide from 44 % to 68 %.

After cassette preparation, restriction digest and purification, the cassettes were ligated into the four Vκ consensus genes using the unique restriction sites *Bbs*I and *Msc*I, and the ligation mixtures were electroporated into *E. coli* TG1 cells. We obtained 6×10^8 independent colonies, and hence an almost complete coverage of the theoretical diversity. The quality of the cassettes was then checked by sequencing 235 independent clones. A total of 175 clones (75 %) were completely correct and showed the library composition as planned. Four clones contained an unplanned amino acid at one position, which was most likely due to single-base mutations introduced during cassette preparation, three clones contained a one-base and six clones contained a one-codon deletion in the trinucleotide-encoded region. All other non-correct clones had the library cassette inserted twice or in the reverse orientation, or they contained one-base deletions in the 5′ mononucleotide region of the oligonucleotide. In order to obtain more statistical data on codon incorporation, all codons originating from trinucleotide positions were analyzed. Figure 7(b) shows the result of that analysis. Over-

all, the data are in excellent agreement with the expected distribution. Q_{L89}, Y_{L91}, and P_{L95} appeared almost exactly as planned at these strongly biased positions.

Vλ CDR3

A total of 147 rearranged human Vλ sequences were collected and analyzed. The lengths of the CDR3s (positions 89 to 96) ranged from seven to 12 residues, the majority (92%) having between eight and ten residues (L3 loop lengths six to eight according to Chothia & Lesk, 1987). Therefore, we decided to construct a CDR library comprising these three different length variants. Analysis of the amino acid composition in the rearranged sequences revealed a high degree of variability at positions 93 to 96, and to a smaller extent at positions 89 to 92. The inspection of antigen contact residues in the case of an antibody of canonical structure 1 (Chothia & Lesk, 1987, see Figure 3) revealed that positions 91, 94, and 96 seem to play the most important role. A single Vλ CDR3 oligonucleotide for all three Vλ consensus genes was designed, where Q_{L89} and S_{L90} were kept constant, since neither position is part of the loop region. Similarly, D_{L92} was fixed as the most frequent amino acid at that position and because its side-chain points away from the binding pocket. Residue 91, which packs against V_H CDR3, was limited to the three most frequent amino acids found in the database (arginine, tryptophan and tyrosine). At positions 93 to 95B, an equimolar mixture of all amino acids except for cysteine and tryptophan was allowed, since cysteine and tryptophan were never found in the rearranged sequences. Position 96 was completely randomized, except for cysteine.

Since the framework 3 region adjacent to the CDR3 of all three Vλ master genes is identical, we could use a single oligonucleotide for all three genes. For oligonucleotide synthesis, three different trinucleotide mixtures had to be prepared comprising three biased codons, 18 or 19 codons (in both cases equally distributed). The three mixtures and their positions in the CDR3 are given in Figure 7(c). On average, the stepwise coupling ratio for trinucleotide mixtures was about 98.9%, the overall yield for the oligonucleotide was 80%. During oligonucleotide synthesis, we used four consecutive sub-stoichiometric coupling steps at the triplet position corresponding to residue 95A. Thereby, we created an oligonucleotide with variable length covering CDR3 lengths between eight and 11 amino acid residues, with the smallest fraction having a CDR3 length of 11 residues. The theoretical diversity of these length variants ranged from 3.3×10^5 (eight residues) to 1.9×10^6 (11 residues).

After cassette preparation, restriction digest and purification, the cassette was ligated into the three Vλ consensus genes using the unique restriction sites *BbsI* and *HpaI*, and the ligation mixtures were electroporated into *E. coli* TG1 cells. We obtained 5.7×10^6 independent colonies.

As described above for Vκ, the quality of the oligonucleotide was checked by sequencing (183 independent clones). Again, about 26% of incorrect sequences could be identified, with errors of the type similar to those found in the Vκ CDR3s. A total of 74% of all clones, however, had completely correct CDR cassettes. The amino acid composition was again in very good agreement with the desired distribution, except for Y_{L91}, which was over-represented at the expense of W_{L91} (see Figure 7(c)). The length distribution was also analyzed: we found that the majority contained a CDR3 length of eight (36%) or nine (42%) residues, the rest had a length of ten (21%) or 11 (2%) residues.

V_H CDR3

For the highly variable V_H CDR3s, all available rearranged sequences were grouped together, irrespective of the individual sub-families. A total of 572 sequences were analyzed. The analysis revealed that only position H101 is stongly biased (toward aspartate in 82% of all cases). This is in agreement with the findings that R_{H94} and D_{H101} form a highly conserved salt-bridge (Searle *et al.*, 1995), and that these two residues are critical for the "kinked base" (Shirai *et al.*, 1996) or "bulged torso" (Morea *et al.*, 1998) structure of the CDR3 loop. D_{H101} was therefore kept constant, although this limits the structural variability to only a subset of CDR-H3 conformations, as other structures are seen in antibodies devoid of the R_{H94}-D_{H101} salt-bridge.

Again, the observed variability corresponds well with the information obtained by inspection of antigen contact residues, showing that positions H95 to H100y seem to play the most important role, whereas H100z is involved to a lesser extent (see the legend to Figure 7 for HCDR3 position nomenclature). Position H102 was found not to be important for antibody/antigen interactions (see Figure 3).

When designing the library cassette, we decided to base the composition of the trinucleotide mixtures for all positions except for H100z and H102 on the overall amino acid composition of the natural heavy chain CDR3s. Positions H100z and H102 were analyzed separately. This resulted in three different codon mixtures, named TH1 (for H95 to H100y), TH2 (for H100z), and TH3 (for H102). The compositions of these mixtures are given in Figure 7(a).

Analysis of the length variability of CDR3 (positions 95 to 102) showed a range between four and 28 residues with a maximum at 13.0. Wu *et al.* (1993) found a mean length of 11.6 residues in their analysis of human antibody sequences. To be able to cover such a broad spectrum of length variants, two separate oligonucleotides were synthesized using the sub-stoichiometric coupling approach to create the shorter library CDR3Ha,

comprising five to 22 residues and the longer library CDR3Hb, comprising nine to 28 residues. Since the two length variants were kept separated during library construction (see below), their use might be adapted to the antigen in question. Moreover, by mixing these two libraries appropriately, it is possible to mimic the natural length diversity. The final yields for oligonucleotides CDR3Ha and CDR3Hb were 68 % and 74 %, respectively, and the sub-stoichiometric coupling rates varied between 35 % and 55 %. Based on these coupling rates, a theoretical length distribution for the two libraries CDR3Ha and CDR3Hb was calculated (see Figure 8).

After cassette preparation, restriction digest and purification, the cassettes were inserted into the scFv libraries already containing the randomized V_L CDR3s described above. We mixed all four Vκ and all three Vλ libraries before HCDR3 insertion, but we kept the V_H consensus genes separate (except V_H1A and V_H1B, which were also mixed). Hence, 24 separate libraries were created (V_H1 to V_H6, each either with four κ or three λ genes, and each either with the short or the long HCDR3 cassette). After electroporation into *E. coli* TG1 cells, we obtained altogether 2.1×10^9 independent colonies.

The quality of the V_H CDR3 s were checked by sequencing 257 clones. In Figure 7(a) the amino acid distributions for the trinucleotide mixtures TH1, TH2, and TH3 are given, showing again an excellent agreement with the calculated and designed frequencies. The sequencing results obtained from both V_H CDR3 length variants revealed that both library types follow a Gaussian length distribution, with the maxima at 9.0 and 16.6 residues (Figure 8). Thus, the actual length distribution was shifted towards shorter lengths when compared to the theoretical length distribution, but the whole range of naturally occurring length variants was covered by the two library variants.

The final library was designated HuCAL1. Altogether, we found the fraction of fully correct library members with CDR1l3 and L3 as designed to be 61 %.

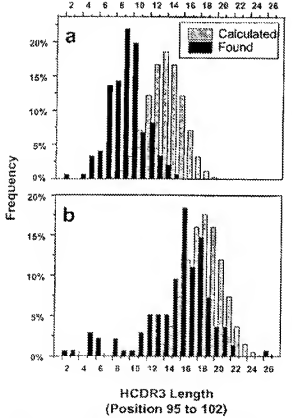

Figure 8. Distribution of CDR3 H length variants in the HuCAL1 libraries. The results from (a) the trinucleotide cassette HCDR3a, and from (b) the cassette HCR3b are shown (black columns) and compared to the length distribution as calculated from the substoichiometric coupling (gray columns). For details, see the text.

Diversity and binding constants

Phage-display as well as ribosome-display selection experiments were performed against a variety of antigens, including proteins, peptides, or whole cells. The HuCAL1 library comprosing all 49 combinations was used for selection experiments. Two or three panning rounds of phage display, or five or six rounds of ribosome display were performed in each case. After the final round, the selected scFv genes were subcloned as a pool in an expression vector and the transformants were screened for binding using ELISA or FACS assays. Details about the selection experiments and the characterization of binders will be given elsewhere (Krebs *et al.*, unpublished results; Hanes *et al.*, unpublished results). In the great majority of cases, many different scFv fragments could be identified, which bound the antigen specifically. The V_H and V_L framework usage for the first 250 specific binders selected from HuCAL1 *via* phage display is given in Table 3. All V_H and V_L frameworks could be selected, and so far 42 of the 49 framework combinations were found to be used. While the V_H4 gene segment is rarely used, the V_H3 gene segment predominates. The predominance of V_H3 occurs also in nature (see Table 1) and is even higher in other libraries (Griffiths *et al.*, 1994; Vaughan *et al.*, 1996). All other HuCAL frameworks seem to be used with similar frequency. There is also a considerable variation in V_H CDR3 length: the first 250 specific binders range from four to 24 residues (data not shown).

Selected binders were purified to homogeneity using affinity chromatography or IMAC, and their monovalent binding constants were measured using surface plasmon resonance (BIAcore). As shown in Table 4, binding constants of peptide binders were in the micromolar range, whereas affi-

Table 3. Framework usage

	κ1	κ2	κ3	κ4	λ1	λ2	λ3	Σ	%
H1A	4	1	3	3	1	3	7	22	9
H1B	3		4	2	1	6	1	17	7
H2	2	5	11	7	11	4	15	55	22
H3	12	5	16	11	15	7	25	91	36
H4					2		1	3	1
H5	4	1	5	3	1	5	14	33	13
H6	8		3	5	2	3	8	29	12
Σ	33	12	42	31	33	28	71	250	
%	13	5	17	12	13	11	28		

For each of the 49 HuCAL framework combinations, the number of specific scFvs from a collection of 250 binders against about 50 different antigens (haptens, peptides, proteins) is shown. All clones have been selected by phage display. The identity of the framework was determined by sequencing.

nities to protein antigens were usually in the low nanomolar range.

Discussion

Here, we describe the realization of the concept of fully synthetic human antibody libraries, designated HuCAL, which are built on seven V_H and seven V_L consensus frameworks, yielding 49 combinations in total.

We have extensively used these first libraries for the successful selection of highly specific binders against all kinds of antigens, including haptens, DNA, peptides, and proteins, including cell-bound receptor antigens (unpublished results). Intrinsic affinities down to the sub-nanomolar range were found against protein antigens, and the majority of binders were found to have dissociation constants between 1 and 1000 nM after only two rounds of selection. All frameworks have been selected, the selected antibodies could be shown to be expressed in good yields, they are suprisingly stable against thermal denaturation, and can be used in typical applications like ELISA, immunoblotting, FACS analysis, immunoprecipitation or immunohistochemistry even without any affinity maturation steps, verifying the successful design of completely synthetic human antibodies described in this study.

Strategy of modular design

The 49 consensus genes were derived by a stepwise analysis of human antibody sequences. First, the collected sequences were grouped into families according to sequence homology. Second, the

Table 4. Affinities of HuCAL scFvs

Antigen	scFv	Framework	Affinity (nM; BIAcore)	$k_{on} \times 10^5$ (M^{-1} s^{-1})	$k_{off} \times 10^{-2}$ (s^{-1})	App. size (kDa; SEC)
ICAM-1[a]	ICAM1-1	H3λ3	9.4	2.13	0.20	32
ICAM-1[a]	ICAM1-15	H5λ2	72.7	1.72	1.25	27
Insulin[b]	C59	H1Aκ1	0.082[e]	-	-	32
Insulin[b]	A21	H3κ2	8[e]	-	-	32
CD11b[c]	Mac1-5	H2λ1	1.0	7.92	0.09	36
CD11b[c]	Mac1-29	H2λ3	1.2	1.76	0.03	25
EGFR (human)	A9-1	H2κ2	246	1.34	3.30	25
Mac1 peptide[d]	3B2	H3λ2	1130	1.85	21.0	30
Hag peptide[d]	C22-2	H3κ4	610	1.41	8.6	28
NFκB peptide[d]	27HA1	H3λ3	1600	0.55	8.8	32

Affinity of FPLC-purified antibody monomers measured by SPR on a BIAcore biosensor. Antigens were coupled to CM5 sensor chips. In order to avoid contamination with multimeric variants, the monomeric scFv fragments were isolated by size-exclusion chromatography (SEC) (Krebs *et al.*, unpublished results; Hanes *et al.*, unpublished results).

[a] The extracellular part of human ICAM-1 (residues 26 to 479) was used as antigen.

[b] Selection against bovine insulin was carried out with ribosome display (Hanes & Plückthun, 1997; Hanes *et al.*, 1998, unpublished results). These antibodies carry additional point mutations created during PCR amplification.

[c] The I-domai of human CD11b (residues 149 to 353) was used as antigen.

[d] The following peptides were synthesized, coupled to a protein carrier and used for antibody selection

Mac1 peptide: NH_2-C-DAPRSEKSRQELNTIASKPPRDHVF-COOH

Hag peptide: NH_2-C-AGPYDVPDYASLRSHH-COOH

NFκB peptide: NH_2-C-LHVTKEKV-COOH

[e] Affinities determined with the inhibition BIAcore method, in which a mass transport-limited on-rate is measured as a function of antigen present in solution (Hanes *et al.*, 1998).

usage for each germline gene was analyzed by calculating for each rearranged sequence in the database the germline gene from which it was derived. Third, the families of frequently used antibody genes were analyzed in terms of structural diversity of the antigen binding loops, following the concept of canonical CDR conformations established by Chothia and co-workers (Chothia *et al.*, 1989). Fourth, consensus sequences were derived from the rearranged sequences, and grouped into families of frequently used human antibodies. Altogether, the analysis resulted in seven V_H, four Vκ and three Vλ consensus sequences, and our analysis suggests that this small set of consensus genes covers almost the entire structural repertoire encoded in those human antibody germline genes that were found to be used during the immune response.

Reducing the human antibody repertoire to 49 distinct Fv frameworks, yet without reducing structural diversity, made it feasible to obtain the sequences *de novo* by gene synthesis, which enabled us to incorporate several features into the genes that facilitate library construction, affinity maturation and *E. coli* gene expression. Moreover, the separate construction of the genes and the resulting libraries allowed detailed analysis of each master framework under defined conditions, which is not possible with antibody phage-display libraries derived from natural sequences by PCR cloning. Particularly, the presence of unique restriction sites across the whole library makes it possible to shuffle CDRs and frameworks, even at the level of pools, and without knowledge of the sequence of the antibodies. Furthermore, the approach is modular and can incorporate future knowledge of antibody structure, folding and stability, as individual framework pieces can easily be replaced in future versions.

The availability of separate libraries for each of the combinations allows one to analyze the performance of separate framework combinations and a direct comparison with results obtained from a mixture or the natural immune response. It also provides a way to force the selection against different epitopes on the same protein, which can be a very crucial feature given that *in vivo* applications may require the blocking of a binding site on a receptor by the antibody, while a different epitope on the receptor may be completely immuno-dominant. In this case, the preferentially selected but unwanted framework combination can simply be left out. Alternatively, separate affinity enrichments with subsets of frameworks can be carried out to enforce the binding of diverse epitopes. In addition, further analysis of the performance of this and other libraries may show that particular framework combinations contribute little to the pool of selected binders, while others need to be provided with more initial diversity in CDR1 and CDR2. The number of frameworks is, of course, arbitrary and can be adjusted by addition of new and subtraction of unnecessary ones.

Expression and folding properties

The HuCAL genes were adapted to *E. coli* codon usage. While we indeed found superior expression behavior from most of the synthetic genes, this probably reflects favorable protein folding properties (see below), although the avoidance of codons used only very rarely is at least a prerequisite for high expression yields. The consensus frameworks described here may be an interesting basis for elucidating the framework contributions to differences in folding yield during recombinant antibody expression and to thermodynamic stability. The absence of large differences in expression behavior between the consensus frameworks may improve library quality, since the probability of clones being eliminated during library selections due to very different effects of distinct antibody sequences on the bacterial cell physiology is minimized.

In this context, it is interesting to note that the high-expressing humanized antibody hu4D5, which was shown to be expressed 10-50-fold better in *E. coli* than the murine parental antibody (Carter *et al.*, 1992a), was designed using human consensus frameworks derived from the subfamilies V_H3 and Vκ1 (Carter *et al.*, 1992b). The human V_H3 germline gene 3-23 (DP-47), which is most homologous (99 % identity) to the HuCAL consensus amino acid sequence of the V_H3 germline subfamily, is also the most frequently used V_H3 germline gene (see Table 1) and it is very frequently found in antibody phage-display libraries based on human genes (Griffiths *et al.*, 1994; Vaughan *et al.*, 1996; Dorsam *et al.*, 1997; Boel *et al.*, 1998; Sheets *et al.*, 1998). Our theoretical analysis (unpublished results) showed that this framework has very few of the recognized sequence problems. Such problem spots include exposed hydrophobic residues that might promote misfolding and aggregation (Nieba *et al.*, 1997), non-Gly residues in positions with conserved positive phi angles, proline in position H40 (Knappik & Plückthun, 1995) and the disruption of the highly conserved charge cluster around R_{H66}/D_{H86} and R_{L61}/D_{L85} (Proba *et al.*, 1998).

It is reasonable therefore, to hypothesize that consensus sequences, which are closely related to phyiogenetically old progenitor genes, are better adapted to folding in an environment like the *E. coli* periplasm, where probably most of the folding catalysts and chaperones, which normally act on the folding pathway in the ER lumen of the antibody producing B-cell, are absent. It is tempting to speculate that a consensus sequence defines a point in sequence space from which the observed sequences have diverged through genetic drift until the function of the protein is no longer maintained. This speculation is supported by experiments (Steipe *et al.*, 1994), where a clear correlation between degree of deviation from the consensus sequence and loss of thermodynamic stability of a murine V_L domain was found. Recent studies (Wörn & Plückthun, 1999) have shown

that, taking all available information into account, very stable and well-expressing antibodies can be engineered, suggesting that the β-sandwich framework is, in principle, a highly stable scaffold. Yet, most antibodies have diverged far from this point, first in the course of gene duplication during evolution of the locus, then in the V(D)J rearrangement where unfavorable CDR3s may be introduced, and finally in the somatic mutations, yielding antibody domains of very marginal biophysical integrity. The mouse repertoire is thought to be significantly larger than the human repertoire (Almagro *et al.*, 1998) and thus more deviations from the optimum are genetically encoded, partially explaining the difficulties in expressing antibody fragments derived from murine hybridomas. Several residues experimentally shown to be non-optimal (Spada *et al.*, 1998; Proba *et al.*, 1998) have been found to be encoded in some of the mouse germline genes, but in none of the human genes. Such residues are totally avoided in the present design.

Affinity maturation

The synthetic HuCAL genes were designed to contain unique restriction sites flanking the regions encoding the antigen binding loops, thereby making all six CDR regions accessible for diversification. The resulting modular gene structure, in combination with pre-built CDR library cassettes, will allow the rapid randomization of each CDR loop. We have constructed trinucleotide-based LCDR1 and HCDR2 cassettes using a design procedure identical with that described here for CDR3 cassettes (unpublished results). Hence an iterative randomization procedure can be envisaged, where the pool of binding sequences obtained after initial library selections can serve as starting material for the next iteration. Such a protocol would mimic the process of affinity maturation by somatic hypermutation observed during the natural immune response, even though the mechanism for achieving this would be different. It may be reasoned that this will be more efficient, as more of the mutations will be targeted to the region of interest. So far, the CDR walking process has been time-consuming, since the protocols and the CDR libraries had to be established for each individual antibody sequence. By using cassettes and the conserved restriction sites of the synthetic genes, however, an optimization of pools is possible, and the procedure is much more convenient. It has been shown now by several groups that the process of CDR walking, i.e. the iterative randomization of CDRs followed by stringent selection protocols, improved binding affinity of distinct antibody sequences dramatically (Yang *et al.*, 1995; Schier *et al.*, 1996b; Barbas & Burton, 1996; Rosok *et al.*, 1998; Wu *et al.*, 1998), and intrinsic affinities in the picomolar range could be obtained by this approach.

Nevertheless, framework residues can have indirect effects on binding by affecting the CDR conformations (Foote & Winter, 1992; Saul & Poljak, 1993), and a complete refinement may have to include these regions as well, e.g. by gene shuffling (Patten *et al.*, 1997) or ribosome display (Hanes *et al.*, 1998). Recently, the latter approach has been applied to the HuCAL1 library, and binders with sub-nanomolar affinities to several antigens have been obtained that do carry further mutations introduced by PCR (unpublished results).

Trinucleotide mixtures for CDR libraries

Using the 49 combined HuCAL frameworks, the initial libraries were created by randomizing two of the six CDR regions using trinucleotide building blocks. Sondek & Shortle (1992) first reported the use of a mixture of two trinucleotide phosphoramidites, but found a coupling yield of only 4% and large differences of relative coupling ratios. Virnekäs *et al.* (1994) showed that coupling of trinucleotide mixtures can be achieved with coupling yields as high as 96-98.5%, by carefully excluding traces of water during preparation of the phosphoramidite mixtures for coupling. However, in a first experiment using eight different trinucleotides, the individual codons were introduced with different frequencies (between one and 15 times within 63 positions being sequenced). No further improvement has been reported by other groups using similar building blocks (Lyttle *et al.*, 1995; Ono *et al.*, 1995; Kayushin *et al.*, 1996). Braunagel & Little (1997) used the trinucleotides described by Kayushin *et al.* (1996) in their approach to create a single-framework antibody library. However, no sequencing results were given to show the quality of the starting library or the distribution of individual codons.

We found that mixtures of trinucleotide phosphoramidites can be coupled in excellent yields. Oligonucleotides with a length of more than 100 bases and containing ten to 15 randomized positions have been successfully synthesized. Furthermore, no bias was found in most cases and trinucleotide-directed mutagenesis appears now to be the method of choice to achieve full control over the variability.

The option of using sub-stoichiometric coupling steps during oligonucleotide synthesis opens up a novel way of creating diversity by sequence and by length variation in a single oligonucleotide. We used sub-stoichiometric coupling for the generation of Vλ and V_H CDR3 libraries, and indeed it was possible to create CDR3s of different length with this method. However, the distribution of different length variants was in all cases shifted to shorter library members than calculated, suggesting that the stepwise coupling yields calculated from measuring the concentration of trityl cations, cleaved off the 5'-end, is higher than the actual coupling yield, i.e. the percentage of oligonucleotide chains being elongated during the sub-stoichio-

metric step. However, the parameters influencing the outcome of the sub-stoichiometric approach have not been studied in detail.

We decided initially to start with a diversification of CDR-L3 and CDR-H3, to imitate natural antibody generation. During the natural process of initial antibody generation, which results from genome rearrangements in the developing B-cell, most of the initial diversification is located in the V_H CDR3 region (VDJ-joining) and, to a lesser extent, the V_L CDR3 (VJ-joining). In the 3D structures of antibodies, both CDR3s form the so-called inner ring of the antigen binding site, and most of the antigen contacts are formed by residues located there (see Figures 1 and 3).

Comparison to semi-synthetic antibody libraries

The use of defined frameworks as the basis for generating an antibody library has been described before. Initial work on randomizing just CDR-H3 (Barbas *et al.*, 1992) has since then been extended to Vκ CDR3 (Barbas *et al.*, 1993; Yang *et al.*, 1995; Söderlind *et al.*,1995) or to single frameworks with all CDRs being randomized (Hayashi *et al.*, 1994; Iba & Kurosawa, 1997). Furthermore, sets of V_H genes extended with PCR primers that encode CDR-H3 libraries have been combined with a single V_L gene (Nissim *et al.*, 1994), or a limited set of V_L genes (De Kruif *et al.*, 1995), or a randomized repertoire of V_L genes (Griffiths *et al.*, 1994).

Most of the semi-synthetic human antibody libraries constructed so far focussed on exclusively randomizing CDR3s. For V_H, in most approaches A_{H89} B_{H94} and D_{H101} were kept constant, and positions H95 to H100a, and usually H102 as well, were randomized (Hoogenboom & Winter, 1992; Barbas *et al.*, 1993, 1994; De Kruif *et al.*, 1995). The length of the CDR3s varied between six and 20 residues, with a preference for loops with six to 14 amino acid residues. De Kruif *et al.* (1995) constructed a set of eight CDRs between eight and 17 residues long, comprising completely and semi-randomized stretches.

For Vκ CDR3, usually residues L92 to L96 were randomized (Barbas *et al.*, 1993, 1994; Yang *et al.*, 1995; Söderlind *et al.*, 1995). The length of the CDR3s varied between seven and ten residues. Similarly, Hayashi *et al.* (1994) randomized 11 residues (including residue L97 of framework 4) of Vλ CDR3 in their approach to construct a one-framework library with all six CDRs being randomized. In contrast, Griffiths *et al.* (1994) used a whole set of 21 Vλ (as well as Vκ) germline genes and added, *via* PCR, specific CDR sequences comprising zero to five randomized codons. In all cases, codons were randomized by using mixtures of mononucleotides during oligonucleotide synthesis.

In our CDR3 design, we had to decide whether to stay close to the encoded variety with a preference for sequences actually found in selected antibodies or whether to follow a more daring approach. While, technically, both approaches are equally feasible, as it would depend only on the types of cassettes used, we opted to first examine CDR3 libraries close to the encoded variety. Even in a loop of this size, many combinations will be non-functional, and we wanted to secure a very high number of initial functional molecules. As library selection technology progresses, e.g. by the use of methods such as ribosome display (Hanes & Plückthun, 1997; Hanes *et al.*, 1998), much larger libraries will be screenable, and a larger set of variants may be simultaneously present, including those with structural defects.

When using the known rearranged sequences as a guide, it becomes an important question to what degree they represent "frozen accidents", explainable only by their evolutionary ancestry both at the germline and somatic level, or whether they are truly positively selected or are even due to genetic hotspots, encoded into the DNA sequence. The processes underlying somatic hypermutation are still not well understood. It was shown that heterologous genes replacing V gene segments undergo hypermutation *in vivo* as well (Yélamos *et al.*, 1995), and therefore it seems very unlikely that the V genes themselves determine at the genetic level where hypermutation occurs. A more reasonable explanation would be that selection determines which mutations finally survive. Various efforts have addressed this question (see, for example, Dorner *et al.*, 1998).

Weighing all arguments, we decided to take the natural distribution as our starting point. The modular approach permits any desired optimization strategy to be readily be carried out, once primary binders have been obtained, such as the introduction of V_L CDR1 and/or V_H CDR2 cassettes into single binders, or even pools of binders, since the sequences share identical restriction endonuclease sites adjacent to the CDRs. It would be also easily possible, for example, to keep the CDR3s of the selected pool of primary binders constant and shuffle V_H frameworks with randomized CDR2s. Alternatively, new sets of CDR3 libraries can be designed based on sequence motifs identified in the pool of primary binders. Furthermore, chain shuffling or even shuffling of elements such as CDRs or frameworks can now performed by restriction digest and religation.

Since HuCAL is fully synthetic, it is always possible to control the individual steps by analyzing the restriction pattern of individual clones or by sequencing, with artifacts being easily identified, whereas an immune repertoire cloned *via* PCR is more or less a black box.

By these means, searching the sequence space of human antibodies will be much faster and more efficient than by using the conventional approaches. Finally, we expect that the careful analysis of selected sequences will contain a wealth of structural information that can flow into subsequent versions of the library.

Conclusions and perspective

The HuCAL concept is based on covering the essential features of the human antibody repertoire with a minimal number of different sequences, which are designed to facilitate extensive manipulation with standard protein engineering techniques. The 49 combinations of master genes have been cloned as scFv genes in both orientations and as F_{ab} genes. Other formats like Fv fragments stabilized for example by disulfide-bridges (Glockshuber *et al.*, 1990; Brinkmann *et al.*, 1995; Rodrigues *et al.*, 1995) or fragments without any disulfide bonds (Wörn & Plückthun, 1998) useful for intrabody approaches (Cattaneo & Biocca, 1999; Wörn *et al.*, 2000) are easily adaptable and can be analyzed on the level of the master genes before actual library generation. Libraries can be rapidly created by inserting pre-built CDR cassettes into each of the 49 genes either separately or as mixed sequence pool, and the analysis of binding variants is facilitated by the fact that only small regions in the sequence are varied and that the three-dimensional models of all master frameworks have been built. It may therefore be possible for the first time to investigate experimentally why nature has evolved the distinct structural motifs found in the human antibody repertoire, and whether there are correlations of antibody structure with antigen class, antibody affinity and specificity. Future versions of HuCAL may therefore be enriched with antigen-type specific features.

Materials and Methods

Bacterial strains, phages, vectors

Molecular cloning was carried out using the *E. coli* strains JM83 (Yanisch-Perron *et al.*, 1985), XL1-Blue (Stratagene) or Top10 (Invitrogen). For expression experiments, JM83 was used. Phage-display libraries were generated and propagated using *E. coli* TG1 as host strain and M13K07 or VCSM13 as helper phage (all from Stratagene). The products from gene synthesis were cloned in pZero-1 (Invitrogen) or pCR-Script SK(+) (Stratagene) for sequencing. The pBS vector series used for antibody cloning and for expression analysis is a derivative of the phage-display vector pAK100 (Krebber *et al.*, 1997). The vector pBS10 contains the mature *pIII* gene preceded by a region encoding the *ompA* signal sequence, a FLAG tag and an *EcoRI* cloning site between the *XbaI/HindIII* cloning sites of pAK100. The pBS10 vector was modified as follows in order to allow assembly of synthetic antibody genes. First, an oligonucleotide cassette encoding a synthetic *phoA* signal sequence (created by annealing the oligonucleotides O5phoA and O3phoA; all oligonucleotides constructed during this work are given in Table 2 of the Supplementary Material) was inserted into the *XbaI/EcoRI* sites. The resulting construct was designated pBS11. This *phoA* gene fragment contained a unique *SapI* site, which was later used for insertion of V_H genes for the generation of F_{ab} fragments. The *phoA* gene fragment was extended by inserting a cassette created by annealing the oligonucleotides O5phoA_F and O3phoA_F into pBS11 *via SapI/EcoRI*, thereby introducing the short improved FLAG tag (DYKDE; Knappik & Plückthun, 1994). The resulting vector, designated pBS12, was later used for the assembly of scFv genes in the H-L orientation as well as for expression analysis. Second, the *XbaI/EcoRI* fragment from pBS10 was replaced by a cassette created by annealing the oligonucleotides O5stII and O3stII, thereby introducing a *stII* signal sequence containing a unique *NsiI* site, which was later used for cloning of V_L genes for the generation of F_{ab} fragments. The resulting vector was designated pBS13. The *stII* gene fragment was extended by inserting a cassette created by annealing the oligonucleotides O5stII_F and O3stII_F into pBS13 *via NsiI/EcoRI*, thereby introducing the short improved FLAG tag. The resulting vector, designated pBS14, was later used for the assembly of scFv genes in the L-H orientation. Vector pBS13b was constructed by removing the *MssI* site in the *cat* resistance marker gene. The phage display vector pIG10.3 is a derivative of pIG10 (Ge *et al.*, 1995), where the first 249 codons of the mature full-length gene III were deleted. Briefly, the *EcoRI/HindIII* restriction fragment in the phagemid pIG10 was replaced by the c-myc tag for detection with the monoclonal antibody 9E10 (Munro & Pelham, 1986) followed by an amber codon and the truncated version of the gene III through PCR mutagenesis. The construction of the pMorph vector series, which is compatible with the HuCAL restriction sites and which was used for library cloning, will be described elsewhere (unpublished results). All vectors were constructed using site-directed mutagenesis (Kunkel, 1985), recursive PCR (Prodromou & Pearl, 1992) and overlap-extension PCR (Ge & Rudolph, 1997), and all constructs were subsequently verified by DNA sequencing (SequiServe, Vaterstetten, Germany).

Collection of human antibody sequences

Functional human germline sequences were downloaded from Genbank (Benson *et al.*, 1997), from the Kabat database† and from Vbase‡. Rearranged sequences were downloaded from Genbank and from the Kabat database. Kabat dump files were downloaded, variable domain amino acid sequences extracted and converted to the one-letter code. Sequences less than 90% complete or containing multiple undetermined residues in the regions of interest were eliminated. The automatic alignment generated by the program Pileup (Wisconsin Package, Version 8.1, 1995, Genetics Computer Group, Madison, WI, USA) was manually corrected to shift the gaps to the closest positions where they could be accommodated in the three-dimensional structure. The sequence files were converted and imported into Microsoft Excel®, where all subsequent alignments and analysis procedures took place. All alignments, numbering and loop regions (L1-L3, H1-H3) are according to structural criteria defined by Chothia and colleagues (see Chothia *et al.*, 1992; Tomlinson *et al.*, 1995; Williams *et al.*, 1996). CDRs were labeled as described by Kabat *et al.* (1991), even though this does not always correspond to the structural definition. Amino acid sequences are given in the single letter code according to standard IUPAC nomenclature. Germline sequences are named according to accepted locus nomenclature for each segment (Giudicelli *et al.*, 1997).

† ftp://ftwis.bme.nwu.edu/pub/database
‡ http://www.mrc-cpe.cam.ac.uk/imt-doc

Statistical analysis of the coverage of sequence space

After alignment and numbering according to Kabat, the databases were normalized by checking for multiple entries of closely related sequences, which we thought would indicate an artificial bias towards a specific set of rearranged sequences. Subsequently, the rearranged and the germline sequences were grouped into the various subfamilies. To assign the nearest germline to each rearranged sequence, the number of identities of a given rearranged sequence to each germline sequence was scored from position 1 to 92 (V_H) or position 1 to 95 (Vκ) or 1 to 958 (Vλ). If the result was ambiguous, e.g. the rearranged sequence was equidistant from two or more germline sequences, or if the best hit gave less than 80% identity, indicating either a very high level of somatic mutations or the origin from an at the time unknown germline gene, the rearranged sequence was omitted from the analysis.

By this analysis, the subfamilies that are used frequently by the human immune system were identified. The databases of rearranged sequences were used to calculate a consensus sequence for each frequently used subfamily. This was done by counting the number of amino acid residues used at each position (position variability) and subsequently identifying the amino acid residue most frequently used at each position. The consensus sequences were cross-checked with the consensus of the germline families to see whether the rearranged sequences were biased at certain positions towards amino acid residues that do not occur in the collected germline sequences, but this was found not to be the case. Subsequently, the CDR1 and CDR2 regions of the consensus sequences were replaced with the corresponding regions of the germline sequences that were most frequently used by the human immune system. For the framework 4 region, the consensus of all rearranged sequences was chosen. For each of these consensus sequences, the most homologous rearranged sequences were then identified and used for validating the consensus by identifying all framework residues that differed between the consensus and the most homologous rearranged sequences. These residues were regarded as artificial and checked by two means: first, the local context of the artificial residue was compared with the corresponding stretch of all the rearranged sequences in the database; and second, the long-range interactions of amino acid residues at these positions were analysed. To this end, the structures of human antibodies available from the Brookhaven Protein Database were analyzed, and the contacts of all side-chains were tabulated. If a certain artificial residue in the consensus sequence was found in the local context of rearranged sequences, and if this residue was not involved in side-chain interactions according to the structural analysis, it was kept at this position. Otherwise, the next most common residue was chosen and analyzed as described above. Finally, the consensus sequences were compared to the corresponding germline sequences and the number of differences were tabulated.

Molecular modeling

The structures of the HuCAL domains were predicted by homology modeling using the Homology, Biopolymer and Discover modules of the program Insight II version 95 (Biosym/MSI, San Diego, CA). To align different templates for the comparison of their conformation, a least-squares fit of the C^{α}-positions of residues H2–H7, H19–H23, H34–H40 (gapped according to structural criteria, not according to Kabat), H44–H50, H67–H71, H78–H82, H88–H94 and H102–H108 (V_H) or L3–L7, L20–L24, L33–L39, L43–L49, L62–L66, L71–L75, L84–L90 and L97–L103 (V_L) was performed. The experimental structures displaying the highest degree of sequence similarity to the different HuCAL constructs are listed in Table 1 of the Supplementary Material. Structural differences between these templates were analyzed to identify the sequence differences responsible for the deviations. The conformation of the dummy CDR3 s was taken from the structure of the humanized 4D5 version 8 (PDB entry 1FVC). Coordinates were assigned using the Homology module and the resulting models checked for steric clashes and cavities before energy minimization (module Discover, CFF91 forcefields). The stereochemical quality of the final domain models was evaluated with the program PROCHECK† (Laskowski *et al.*, 1993; Morris *et al.*, 1992).

Gene synthesis and assembly

Consensus amino acid sequences were back-translated into DNA sequences using the GCG software package (Genetics Computer Group, Madison, WI, USA) and a Codon definition file that included only the codons that are used frequently in *E. coli*‡. All possible silent (and commercially available) restriction sites based on version 501 of the REBASE§ list of restriction enzymes (Roberts & Macelis, 1998) were subsequently identified in the resulting DNA sequences and tabulated. These tables were used to identify all cleavage sites that were located close to the CDR and framework borders, and that could be introduced into all genes of the three classes (V_H, Vκ or Vλ) simultaneously at the same position. Further editing was done as described in Results. For each of the 14 resulting genes, six overlapping oligonucleotides were designed. Since both the CDR3 and the framework 4 gene segments were identical in all Vκ, Vλ and V_H genes, respectively, this part was constructed only once in each case. The region of overlap was chosen to give a theoretical t_m of 58°C (corresponding to a ΔG of about −20 kcal/mol), and the 3′ nucleotide was chosen to be either C or G. The design was examined and optimized in terms of potential stem-loop formation, dimer formation and potential unspecific hybridization sites with all other oligonucleotides (duplex formation) using the VectorNTI® software (Informax, Inc.). PCR assembly (Prodromou & Pearl, 1992) was performed by mixing 200 pmol of each of the oligonucleotides in a 100 µl reaction volume containing 20 nmol of dNTPs and five units of Pfu polymerase (Stratagene). After a first cycle with three minutes at 94°C, two minutes at 60°C and one minute at 72°C using a hotstart procedure, 31 PCR cycles were performed (one minute at 94°C, two minutes at 60°C and one minute at 72°C), the products were purified using the QIAgen PCR purification kit and blunt-end ligated with either the pCR-Script KS(+) (cut with *Srf*I) or the pZero-1 vector (cut with *Eco*RV).

† www.biochem.ucl.ac.uk/~roman/procheck/procheck.html

‡ ftp://ftp.ebi.ac.uk/pub/databases/codonusage/eco.cod

§ http://www.neb.com/ftp_info/rebase/

Insert containing clones were screened by blue-white selection (pCR-Script KS(+)) or directly picked (pZero-1) and sequenced.

The seven synthesized V_H genes covered the sequences from the first unique 5′ restriction site located in the *phoA* signal sequence region (*SapI*) to the last unique 3′ restriction site located in the framework 3 region prior to the CDR3 (*BssHII*). All genes were synthesized with their authentic N termini and without the short FLAG sequence, which was added later during the construction of scFv display vectors. The heavy chain C_H1 domain (subtype IgG1, Genbank accession number A49444) including the V_H framework 4 region was assembled using eight oligonucleotides (OCH1 to OCH8) and inserted into pCR-Script KS(+). The C_H1 gene sequence was designed for optimal *E. coli* codon usage. Additionally, restriction sites for *SalI* and *EcoRI* were incorporated at the 5′- and 3′-ends, respectively, and most internal restriction sites were removed during the gene design. In a second step, the V_H dummy CDR3 region was inserted as a *PstI/StyI* cassette using the oligonucleotides OHCDR3F and OHCDR3 M. The V_H gene fragments (*SapI/BssHII*) were assembled with the CDR3-framework 4-C_H1 sequence (*BssHII/EcoRI*) by a three-fragment ligation with the vector pBS11 (*SapI/EcoRI*), yielding seven Fd fragments for construction of F_{ab} expression vectors. The N-terminal FLAG tag was added later for scFv constructions by cloning the Fd fragments into the vector pBS12 using the restriction enzymes *MfeI* and *EcoRI*.

The four synthesized V_L kappa genes covered the sequences from the unique 5′ restriction site located in the *stII* signal sequence region (*NsiI*) to the unique 3′ restriction site located in the framework 3 region prior to the CDR3 (*Eco57I*). The human kappa constant domain Cκ (Genbank accession number P01834) including the Vκ framework 4 region, the V_L dummy CDR3, and part of the Vκ framework 3 region (including the *Eco57I* restriction site) was synthesized using eight oligonucleotides (OCLK1 to OCLK8) and inserted into pCR-Script KS(+). The Cκ gene sequence was optimized for *E. coli* codon usage, the internal restriction sites except *AscI* were removed and the restriction sites for *BsiWI* and *StuI* were incorporated at the 5′ and 3′-end, respectively. The Vκ gene fragments (*NsiI/Eco57I*) were then assembled with the CDR3-framework 4-Cκ sequence (*SphI/Eco57I*) by a three-fragment ligation with the vector pBS13 (*SphI/NsiI*), yielding four kappa light chain fragments for construction of F_{ab} expression vectors.

The three synthesized V_L lambda genes covered the sequences from the unique 5′ restriction site located in the *stII* signal sequence region (*NsiI*) to the unique 3′ restriction site located in the framework 3 region prior to the CDR3 (*BbsI*). All genes were synthesized using their authentic N termini, i.e. without the aspartate-isoleucine stretch encoded by an *EcoRV* site used for the Vκ genes. The human lambda constant domain Cλ-1 (Genbank accession number P01842) including the Vλ framework 4 region, the V_L dummy CDR3, and part of the Vλ framework 3 region was assembled as *BbsI/SphI* fragment by complete gene synthesis with 12 oligonucleotides. The Vλ gene fragments (*NsiI/BbsI*) were assembled with the CDR3-framework 4-Cλ sequence (*BbsI/SphI*) by a three-fragment ligation with the vector pBS13b (*SphI/NsiI*), yielding three lambda light chain fragments for construction of F_{ab} expression vectors. In order to assemble V_L-V_H scFv vectors, the Vλ gene fragments were further modified. the Vλ gene fragments were PCR amplified from pBS13b using the forward primers OLxFwIDIP (where x denotes the Vλ sub-family) and the backward primer OLFw4 M, and the PCR products were blunt-end ligated into pBS14_scH3 (a V_L-V_H scFv expression vector constructed as described below), which had been cut with *EcoRV/BsiWI* and made blunt-ended by teatment with S_1 nuclease. The resulting three plasmids were named pBS14_scλ$_{1-3}$H3 and contained Vλ genes where the two N-terminal codons had been changed to the *EcoRV* recognition sequence encoding aspartate-isoleucine, in order to allow the same scFv cloning protocol as used for the Vκ genes. These plasmids were used for assembly of the V_L-V_H scFv expression vectors (see below). In order to assemble V_H-V_L scFv vectors, a cassette constructed by annealing the oligonucleotides OLEco5 and OLEco3 was inserted into the lambda light chain containing vectors pBS13b_Vλ$_{1-3}$Cλ cut with *MscI/EcoRI*, thereby replacing the CL constant domain gene fragment by an in-frame *EcoRI* site. The resulting three plasmids were designated pBS13b_Vλ$_{(1-3)}$_E. After cutting these vectors with *XmaI/HindIII*, the 3′-region of each the three Vλ genes including the in-frame *EcoRI* sequence were isolated and inserted into the corresponding pBS14_scλ$_{(1-3)}$H3 vectors, thereby adding the 5′ *EcoRV* and the 3′ *EcoRI* sites to each of the three Vλ genes. These genes were used to assemble the V_H-V_L scFv expression vectors (see below).

F_{ab} expression plasmids were constructed by combining each of the heavy chain Fd fragments cut with *SphI/EcoRI* and each of the light chain fragments cut with *XbaI/SphI* with the pBS13 vector cut with *XbaI/EcoRI* in a three-fragment ligation reaction. The 49 resulting plasmids were verified by restriction enzyme digestions. Here, the Vλ gene fragments contain their authentic N termini, and there is no FLAG tag sequence attached to the antibody F_{ab} genes.

The scFv expression plasmids in the orientation V_L-V_H were constructed as follows: the Cκ gene fragment from pB13_Vκ2Cκ was removed by cutting the plasmid with *BsiWI/SphI* and replaced by an oligonucleotide cassette encoding a 20 amino acid residue linker plus the additional restriction sites *MfeI* and *EcoRI* for later insertion of the V_H genes. The cassette was constructed by annealing the oligonucleotides OLHLiP and OLHLiM. Subsequently, the remaining Vκ and Vλ genes were inserted as *XbaI/BsiWI* fragments and the V_H genes were inserted as *MfeI/EcoRI* fragments.

The 49 scFv expression plasmids in the orientation V_H-V_L were constructed as follows: the CH1 gene fragment from pBS12_VH3CH1 was removed by cutting the plasmid with *BlpI/EcoRI* and replaced by an oligonucleotide cassette encoding a 20 amino acid residue linker plus the additional restriction site *EcoRV* for later insertion of the V_L genes. The cassette was constructed by annealing the oligonucleotides OHLLiP and OHLLiM. Subsequently the Vκ and Vλ genes were inserted as *EcoRV/EcoRI* fragments and the V_H genes were inserted as *XbaI/BlpI* fragments. These 49 vectors were used for expression analysis, and the scFv genes were later used for library construction.

Expression analysis

Growth curves and expression data were obtained essentially as described (Knappik & Plückthun, 1995). Briefly, *E. coli* JM83 cultures containing the appropriate

scFv expression vectors were grown at 30°C and induced with 1 mM IPTG. After two hours of expression, cells were harvested, normalized to identical absorbance, lysed and separated into soluble and insoluble cell fractions by centrifugation. The fractions were assayed by reducing SDS-PAGE, blotting and immunostaining using the anti-FLAG antibody M1 (Sigma), and the amount of soluble and insoluble scFv protein produced was quantified densitometrically. The scFv gene H3c2 was used as internal control in each expression experiment.

Expression kinetics were measured as follows: *E. coli* JM83 cells were transformed with the scFv genes cloned in the expression vector pMorph7_FS (unpublished results) and grown in 1 l shaking-flask cultures at 30°C. After induction with 1 mM IPTG, 50 ml of culture was harvested each hour, the cells were normalized to $A = 50$, lysed by sonification, and the crude extracts were stored at −20°C. After ten hours induction, the remaining culture (500 ml) was harvested, lysed, the scFv fragment was purified using Poros Streptactin affinity chromatography (IBA, Göttingen, Germany), and the amount of purified scFv was determined. The functional scFv expression yield at the different time-points was then determined by ELISA measurements, where the purified antibody fragment of known concentration served as internal standard used to calculate the scFv amount based on the ELISA signal obtained.

CDR analysis and library design

The aligned collections of rearranged human antibody V_H and V_L sequences were used for analysis of CDR3 length and composition. For analysis of V_H CDR3 s, all sequences were grouped together because sequence alignments are not possible in this highly diverse region. For Vκ and Vλ CDR3s, the subfamilies were analyzed separately. Within the individual alignments, the CDRs were grouped according to CDR length. Assignment of the individual groups to canonical structures was done according to the rules described by Chothia *et al.* (1989) All analysis was done using Microsoft Excel[9].

Synthesis of trinucleotide-containing oligonucleotides

Synthesis of *O*-methyl trinucleotide phosphoramidites and their application in automated DNA synthesis has been described (Virnekäs *et al.*, 1994). Trinucleotide mixtures were prepared by mixing appropriate stoichiometric amounts of solid phosphoramidites, assuming equal reactivities of all 20 trinucleotides. The mixtures were dried under argon and dissolved to yield 0.1 M solutions as described (Virnekäs *et al.*, 1994). Automated synthesis was performed on an Applied Biosystems DNA synthesizer 380B. The synthesis reagents were obtained from Applied Biosystems and MWG (Ebersberg, Germany). All trinucleotide-based syntheses were performed on columns with polystyrene support, 1000 Å, 40 nmol (Applied Biosystems, art. 401072 and 401075). For synthesizing stretches with mononucleotide building blocks of the oligonucleotides, conventional mononucleotide *O*-cyanoethyl phosphoramidites, and the standard synthesis cycle 5SCEAF (single coupling, 15 seconds wait step) were used. When coupling trinucleotide mixtures stoichiometrically, the standard cycle was changed to double couple, including a 100 seconds wait step after the first, and a 400 seconds wait step after the second coupling. For sub-stoichiometric couplings, the time for delivering activated phosphoramidite solution to the column was reduced to achieve approximately 50% coupling yield. If substoichiometric coupling rates were much higher or lower than 50%, either the time was adjusted for the subsequent couplings to obtain an average yield of 50% over all substoichiometric couplings or an additional substoichiometric coupling step was added. Deprotection of the oligonucleotides was performed as described (Virnekäs *et al.*, 1994). All trinucleotide-containing oligonucleotides synthesized for CDR3 library generation are given in Table 3 of the Supplementary Material.

Cassette preparation

The oligonucleotides were resuspended in TE buffer and purified with an 5200 column (Pharmacia) according to the supplier's manual. The complementary strand was synthesized with Klenow polymerase (New England Biolabs). Approximately 5 nmol of oligonucleotide was mixed with a cassette-specific corresponding primer at a ratio of 1:1.2, respectively, heated for ten minutes to 80°C, followed by slowly cooling to room temperature. 10 µl of a 10 mM dNTP mixture, 15 µl of Klenow buffer, 2 µl of Klenow polymerase and water to 150 µl final volume were added. The fill-in reaction was performed at 37°C for two hours and purified with a Nick Spin column according to the supplier's manual (Pharmacia Biotech). The fill-in reaction was checked by an analytical FMC agarose gel (Biomol). To amplify the fill-in products, PCR reactions were performed using 1 µl of the fill-in reaction mixtures (approximately 25 pmol) and 100 pmol of each primer (fill-in primer plus second cassette-specific primer) in each case (30 cycles, one minute at 94°C; one minute at 56°C, one minute at 72°C). The PCR mixtures were purified with a Nick Spin column. The oligonucleotide library cassettes were prepared for ligation by adding 30 µl TE buffer to 100 µl of the purified PCR product, 150 units of each of the corresponding restriction enzymes, and water to a final volume of 300 µl, and by digesting overnight at 37°C. The cassettes were purified on 4% FMC agarose gels (Biomol), and recovered from the gel *via* BIOTRAP elution (Schleicher & Schuell, Germany) according to the supplier's manual (approximately two hours at 100 V/50-70 mA). The solutions containing the cassettes were desalted with Nick spin columns. The quality of the cassettes was checked by analytical FMC agarose gels (4%).

Generation of the HuCAL1 library

Template V_H vectors were created by inserting the seven HuCAL V_H master genes as Fd genes from the vector pBS13 into the display vector pMorph7 (unpublished results). The V_H CDR3 sequences were then replaced by a 1220 bp dummy fragment containing the β-lactamase gene, thereby facilitating subsequent steps for vector fragment preparation. The template V_H vectors were cut with *SfuI/HindIII* to remove the CH1 gene fragment, and the vector fragments were purified. At this step, the two template vector fragments encoding the V_H1A and V_H1B master genes were mixed in an equimolar ratio, resulting in six V_H vector templates.

Template V_L vectors were constructed by first inserting the HuCAL scFv master genes containing the

HuCAL H3 gene in combination with all seven V_L genes into the vector pMorph4, and second, replacing the V_L CDR3 sequences by the β-lactamase gene dummy sequence as described above. The resulting seven template V_L vectors were then purified, the 1220 bp dummy fragment was removed by cutting with *BlnI/MscI* for Vκ gene-containing vectors and *BlnI/HpaI* for the Vλ gene-containing vectors. The prepared trinucleotide cassettes encoding the V_L CDR3 libraries were then ligated separately with the seven V_L template vector fragments (25 fmol of each vector was ligated with 250 fmol of each cassette for three hours at room temperature), and the ligation mixtures were electroporated in 0.9 ml of *E. coli* TG1 cells, yielding altogether 1.14×10^7 independent colonies. The colonies were scraped off the selection plates, and the V_L CDR3 libraries were stored in 20% (w/v) glycerol at −80 °C. Phagemid DNA of the four Vκ and the three Vλ libraries was prepared and two pools were created by mixing the four Vκ (κ_{mix}) and the three Vλ (λ_{mix}) libraries in an equimolar ratio. The two DNA pools were treated with *StyI/HindIII*, the V_L gene libraries were purified using agarose gel electrophoresis and 75 fmol of each pool was ligated with 23 fmol of each of the six V_H template vectors (see above), and electroporated in 0.3 ml of *E. coli* TG1 cells, resulting in altogether 2.3×10^7 colonies for the 12 library pools. In the final step, these 12 libraries were prepared as DNA, cut with *BssHII/StyI* to remove the β-lactamase dummy gene inside the V_H CDR3 region, and the two V_H CDR3 trinucleotide library cassettes (HCDR3a and HCDR3b) were inserted separately by ligation using the same conditions as above. After electroporation into 7.2 ml of *E coli* TG1 (12 electroporations for each library), we obtained altogether 2.1×10^9 independent colonies. The diversity for each pool was between 0.6×10^9 ($V_H2\lambda_{mix}$) and 3.9×10^9 ($V_H6\kappa_{mix}$). The colonies were scraped off the selection plates, and the 24 HuCAL1 library were stored as aliquots in 20% glycerol at −80 °C.

Data Bank accession numbers

The coordinates of the 14 framework models have been deposited in the RCSB Protein Data Bank, entries 1DGX (Vκ1), 1DH4 (Vκ2), 1DH5 (Vκ3), 1DH6 (Vκ4), 1DH7 (Vλ1), 1DH8 (Vλ2), 1DH9 (Vλ3), 1DHA (VH1A), 1DHO (VH1B), 1DHQ (VH2), 1DHU (VH3), 1DHV (VH4), 1DHW (VH5) and 1DHZ (VH6).

Acknowledgments

We thank Ilona Meyer, Peter Rudolph, Martina Mayer, Isabel Rapp, Ursi Holzinger and Siegfried Hirler for excellent technical assistance, Leodevico Ilag for advice in structural analysis, and Titus Kretzschmar and William Reisdorf for critical reading of the manuscript. We thank Josef Hanes, Barbara Krebs, Titus Kretzschmar, Ralf Ostendorp, Josef Prassler, Christine Rothe, Silke Reiffert, Christiane Schaffitzel and Robert Schier for providing the data given in Table 4. We thank David Fischer, Pharmacia & Upjohn, for supplying us with purified Mac-1 and ICAM-1.

References

Al-Lazikani, B., Lesk, A. M. & Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* 273, 927–948.

Almagro, J. C., Hernandez, I., Ramirez, M. C. & Vargas-Madrazo, E. (1998). Structural differences between the repertoires of mouse and human germline genes and their evolutionary implications. *Immunogenetics*, 47, 355–363.

Baca, M., Presta, L. G., O'Connor, S. J. & Wells, J. A. (1997). Antibody humanization using monovalent phage display. *J. Biol. Chem.* 272, 10678–10684.

Barbas, C. F. & Burton, D. R. (1996). Selection and evolution of high-affinity human anti-viral antibodies. *Trends Biotech.* 14, 230–234.

Barbas, C. F., Bain, J. D., Hoekstra, D. M. & Lerner, R. A. (1992). Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proc. Natl Acad. Sci. USA*, 89, 4457–4461.

Barbas, C. F., Amberg, W., Simoncsits, A., Jones, T. M. & Lerner, R. A. (1993). Selection of human anti-hapten antibodies from semi-synthetic libraries. *Gene*, 137, 57–62.

Barbas, S. M., Ghazal, P., Barbas, C. F., III & Burton, D. R. (1994). Recognition of DNA by synthetic antibodies. *J. Am. Chem. Soc.* 116, 2161–2162.

Barbie, V. & Lefranc, M. P. (1998). The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. *Exp. Clin. Immunogenet.* 15, 171–183.

Barre, S., Greenberg, A. S., Flajnik, M. F. & Chothia, C. (1994). Structural conservation of hypervariable regions in immunoglobulins evolution. *Nature Struct. Biol.* 1, 915–920.

Baskin, B., Islam, K. B. & Smith, C. I. (1998). Characterization of the CDR3 region of rearranged alpha heavy chain genes in human fetal liver. *Clin. Exp. Immunol.* 112, 44–47.

Benson, D. A., Boguski, M. S., Lipman, D. J. & Ostell, J. (1997). Genbank. *Nucl Acids Res.* 25, 1–6.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542.

Boel, E., Boolsma, E., de Kruif, J., Jensze, M., Klingman, K. L., Vandijk, H. & Logtenberg, T. (1998). Phage antibodies obtained by competitive selection on complement-resistant *Moraxella* (*Branhamella*) *catarrhalis* recognize the high-molecular-weight outer membrane protein. *Infect. Immun.* 66, 83–88.

Bothmann, H. & Pluckthun, A. (1998). Selection for a periplasmic factor improving phage display and functional periplasmic expression. *Nature Biotech.* 16, 376–380.

Braunagel, M. & Little, M. (1997). Construction of a semisynthetic antibody library using trinucleotide oligos. *Nucl. Acids Res.* 25, 4690–4691.

Brinkmann, U., Chowdhury, P. S., Roscoe, D. M. & Pastan, I. (1995). Phage display of disulfide-stabilized Fv fragments. *J. Immunol. Methods*, 182, 41–50.

Carter, P. & Merchant, A. M. (1997). Engineering antibodies for imaging and therapy. *Curr. Opin. Biotechnol.* 8, 449–454.

Carter, P., Kelley, R. F., Rodrigues, M. L., Snedecor, B., Covarrubias, M., Velligan, M. D., Wong, W. L., Rowland, A. M., Kotts, C. E. & Carver, M. E. (1992a). High level *Escherichia coli* expression and

production of a bivalent humanized antibody fragment. *BioTechnology*, **10**, 163-167.
Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B., Henner, D., Wong, W. L., Rowland, A. M., Kotts, C., Carver, M. E. & Shepard, H. M. (1992b). Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proc. Natl Acad. Sci. USA*, **89**, 4285-4289.
Cattaneo, A. & Biocca, S. (1999). The selection of intracellular antibodies. *Trends Biotechnol.* **17**, 115-121.
Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901-917.
Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877-883.
Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). Structural repertoire of the human VH segments. *J. Mol. Biol.* **227**, 799-817.
Chothia, C., Gelfand, I. & Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.* **278**, 457-479.
Cook, G. P. & Tomlinson, I. M. (1995). The human immunoglobulin V-H repertoire. *Immunol. Today*, **16**, 237-242.
Cox, J. P., Tomlinson, I. M. & Winter, G. (1994). A directory of human germ-line V kappa segments reveals a strong bias in their usage. *Eur. J. Immunol.* **24**, 827-836.
Cwirla, S. E., Balasubramanian, P., Duffin, D. J., Wagstrom, C. R., Gates, C. M., Singer, S. C., Davis, A. M., Tansik, R. L., Mattheakis, L. C., Boytos, C. M., Schatz, P. J., Baccanari, D. P., Wrighton, N. C., Barrett, R. W. & Dower, W. J. (1997). Peptide agonist of the thrombopoietin receptor as potent as the natural cytokine. *Science*, **276**, 1696-1699.
Dall'Acqua, W. & Carter, P. (1998). Antibody engineering. *Curr. Opin. Struct. Biol.* **8**, 443-450.
De Haard, H. J., Van Neer, N., Reurs, A., Hufton, S. E., Roovers, R. C., Henderikx, P., De Bruine, A. P., Arends, J. W. & Hoogenboom, H. R. (1999). A large non-immunized human Fab fragment phage library that permits rapid isolation and kinetic analysis of high affinity antibodies. *J. Biol. Chem.* **274**, 18218-18230.
De Kruif, J., Boel, E. & Logtenberg, T. (1995). Selection and application of human single chain Fv antibody fragments from a semi-synthetic phage antibody display library with designed CDR3 regions. *J. Mol. Biol.* **248**, 97-105.
Deng, S. J., MacKenzie, C. R., Sadowska, J., Michniewicz, J., Young, N. M., Bundle, D. R. & Narang, S. A. (1994). Selection of antibody single-chain variable fragments with improved carbohydrate binding by phage display. *J. Biol. Chem.* **269**, 9533-9538.
Dorner, T., Foster, S. J., Farner, N. L. & Lipsky, P. E. (1998). Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. *Eur. J. Immunol.* **28**, 3384-3396.
Dorsam, H., Rohrbach, P., Kurschner, T., Kipriyanov, S., Renner, S., Braunagel, M., Welschof, M. & Little, M. (1997). Antibodies to steroids from a small human naive IgM library. *FEBS Letters*, **414**, 7-13.
Foote, J. & Winter, G. (1992). Antibody framework residues affecting the conformation of the hypervariable loops. *J. Mol. Biol.* **224**, 487-499.
Forsberg, G., Forsgren, M., Jaki, M., Norin, M., Sterky, C., Enhörning, A., Larsson, K., Ericsson, M. & Björk, P. (1997). Identification of framework residues in a secreted recombinant antibody fragment that control production level and localization in *Escherichia coli*. *J. Biol. Chem.* **272**, 12430-12436.
Ge, L. & Rudolph, P. (1997). Simultaneous introduction of multiple mutations using overlap extension PCR. *Biotechniques*, **22**, 28 ff.
Ge, L., Knappik, A., Pack, P., Freund, C. & Plückthun, A. (1995). Expressing antibodies in *Escherichia coli* In *Antibody Engineering* (Borrebaeck, C. A. K., ed.), pp. 229-266, Oxford University Press, Oxford.
Georgiou, G., Stathopoulos, C., Daugherty, P. S., Nayak, A. R., Iverson, B. L. & Curtiss, R. (1997). Display of heterologous proteins on the surface of microorganisms-from the screening of combinatorial libraries to live recombinant vaccines. *Nature Biotech.* **15**, 29-34.
Giudicelli, V., Chaume, D., Bodmer, J., Muller, W., Busin, C., Marsh, S., Bontrop, R., Marc, L., Malik, A. & Lefranc, M. P. (1997). IMGT, the International ImMunoGeneTics Database. *Nucl. Acids Res.* **25**, 206-211.
Glockshuber, R., Malia, M., Pfitzinger, I. & Pluckthun, A. (1990). A comparison of strategies to stabilize Immunoglobulin Fv-fragments. *Biochemistry*, **29**, 1362-1367.
Green, N. S., Lin, M. M. & Scharff, M. D. (1998). Somatic hypermutation of antibody genes: a hot spot warms up. *Bioessays*, **20**, 227-234.
Griffiths, A. D., Williams, S. C., Hartley, O., Tomlinson, I. M., Waterhouse, P., Crosby, W. L., Kontermann, R. E., Jones, P. T., Low, N. M., Allison, T. J., Prospero, T. D., Hoogenboom, H. R., Nissim, A., Cox, J. P. & Harrison, J. L., *et al.* (1994). Isolation of high affinity human antibodies directly from large synthetic repertoires. *EMBO J.* **13**, 3245-3260.
Hanes, J. & Plückthun, A. (1997). In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl Acad. Sci. USA*, **94**, 4937-4942.
Hanes, J., Jermutus, L., Weber-Bornhauser, S., Bosshard, H. R. & Plückthun, A. (1998). Ribosome display efficiently selects and evolves high-affinity antibodies in vitro from immune libraries. *Proc. Natl Acad. Sci. USA*, **95**, 14130-14135.
Hayashi, N., Welschof, M., Zewe, M., Braunagel, M., Dubel, S., Breitling, F. & Little, M. (1994). Simultaneous mutagenesis of antibody CDR regions by overlap extension and PCR. *Biotechniques*, **17**, 310, 312, 314-315.
He, Y. Y., Stockley, P. G. & Gold, L. (1996). *In vitro* evolution of the DNA binding sites of *Escherichia coli* methionine repressor, MetJ. *J. Mol Biol.* **255**, 55-66.
Hoogenboom, H. R. & Winter, G. (1992). By-passing immunisation. Human antibodies from synthetic repertoires of germline VH gene segments rearranged *in vitro*. *J. Mol. Biol.* **227**, 381-388.
Hoogenboom, H. R., De Bruine, A. P., Hufton, S. E., Hoet, R. M., Arends, J. W. & Roovers, R. C. (1998). Antibody phage display technology and its applications. *Immunotechniques*, **4**, 1-20.
Iba, Y. & Kurosawa, Y. (1997). Comparison of strategies for the construction of libraries of artificial antibodies. *Immunol Cell Biol.* **75**, 217-221.
Ignatovich, O., Tomlinson, I. M., Jones, P. T. & Winter, G. (1997). The creation of diversity in the human

immunoglobulin V-lambda repertoire. *J. Mol. Biol.* 268, 69-77.

Jackson, J. R., Sathe, G., Rosenberg, M. & Sweet, R. (1995). In vitro antibody maturation-improvement of a high affinity neutralizing antibody against IL-1 beta. *J. Immunol.* 154, 3310-3319.

Jirholt, P., Ohlin, M., Borrebaeck, C. A. K. & Söderlind, E. (1998). Exploiting sequence space-shuffling in vivo formed complementarity determining regions into a master framework. *Gene*, 215, 471-476.

Johnson, G., Kabat, E. & Wu, T. T. (1996). Kabat database of sequences of proteins of immunological interest. In *WEIR'S Handbook of Experimental Immunology I. Immunochemistry and Molecular Immunology* (Herzenberg, L. A., Weir, W. M., Herzenberg, L. A. & Blackwell, C., eds), 5th edit. chapt. 6, pp. 6.1-6.21, Blackwell Science Inc. , Cambridge, MA.

Jung, S. & Plückthun, A. (1997). Improving in vivo folding and stability of a single-chain Fv antibody fragment by loop grafting. *Protein Eng.* 10, 959-966.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest*, 5th edit., NIH Publication no. 91-3242, US Department of Health and Human Services, Washington, DC.

Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Schmeits, J. L., Wang, J. & Shimizu, N. (1997). One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res.* 7, 250-261.

Kayushin, A. L., Korosteleva, M. D., Miroshnikov, A. I., Kosch, W., Zubov, D. & Piel, N. (1996). A convenient approach to the synthesis of trinucleotide phosphoramidites-synthons for the generation of oligonucleotide/peptide libraries. *Nucl. Acids Res.* 24, 3748-3755.

Kieke, M. C., Cho, B. K., Boder, E. T., Kranz, D. M. & Wittrup, K. D. (1997). Isolation of anti-T cell receptor scFv mutants by yeast surface display. *Protein Eng.* 10, 1303-1310.

Knappik, A. & Plückthun, A. (1994). An improved affinity tag based on the FLAG peptide for the detection and purification of recombinant antibody fragments. *Biotechniques*, 17, 754-761.

Knappik, A. & Plückthun, A. (1995). Engineered turns of a recombinant antibody improve its *in vivo* folding. *Protein Eng.* 8, 81-89.

Krebber, A., Bornhauser, S., Burmester, J., Honegger, A., Willuda, J., Bosshard, H. R. & Plückthun, A. (1997). Reliable cloning of functional antibody variable domains from hybridomas and spleen cell repertoires employing a reengineered phage display system. *J. Immunol. Methods*, 201, 35-55.

Kunkel, T. A. (1985). Rapid & efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl Acad. Sci. USA*, 82, 488-492.

Langedijk, A. C., Honegger, A., Maat, J., Planta, R. J., van Schaik, R. C. & Plückthun, A. (1998). The nature of antibody heavy chain residue H6 strongly influences the stability of a VH domain lacking the disulfide bridge *J. Mol. Biol.* 283, 95-110.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog.* 26, 283-291.

Lyttle, M. H., Napolitano, E. W., Calio, B. L. & Kauvar, L. M. (1995). Mutagenesis using trinucleotide beta-cyanoethyl phosphoramidites. *Biotechniques*, 19, 274-281.

Matsuda, F. & Honjo, T. (1996). Organization of the human immunoglobulin heavy-chain locus. *Advan. Immunol.* 62, 1-29.

Morea, V., Tramontano, A., Rustici, M., Chothia, C. & Lesk, A. M. (1998) Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J. Mol. Biol.* 275, 269-294.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Struct. Funct. Genet.* 12, 345-364.

Munro, S. & Pelham, H. R. (1986). An Hsp70-like protein in the ER: identity with the 78 kD glucose-regulated protein and immunoglobulin heavy chain binding protein. *Cell*, 46, 291-300.

Nieba, L., Honegger, A., Krebber, C. & Plückthun, A. (1997). Disrupting the hydrophobic patches at the antibody variable/constant domain interface-improved in vivo folding and physical characterization of an engineered scFv fragment. *Protein Eng.* 10, 435-444.

Nissim, A., Hoogenboom, H. R., Tomlinson, I. M., Flynn, G., Midgley, C., Lane, D. & Winter, G (1994). Antibody fragments from a 'single pot' phage display library as immunochemical reagents. *EMBO J.* 13, 692-698.

Oliva, B., Bates, P. A., Querol, E., Aviles, F. X. & Sternberg, M. E. (1998). Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J. Mol. Biol.* 279, 1193-1210.

Ono, A., Matsuda, A., Zhao, J. & Santi, D. V. (1995). The synthesis of blocked triplet-phosphoramidites and their use in mutagenesis. *Nucl. Acids Res.* 23, 4677-4682.

Pallares, N., Frippiat, J. P., Giudicelli, V. & Lefranc, M. P. (1998). The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. *Exp. Clin. Immunogenet.* 15, 8-18.

Patten, P. A., Howard, R. J. & Stemmer, W. P. C. (1997). Applications of DNA shuffling to pharmaceuticals and vaccines. *Curr. Opin. Biotechnol.* 8, 724-733.

Perelson, A. S. (1989). Immune network theory. *Immunol. Rev.* 110, 5-36.

Pini, A., Viti, F., Santucci, A., Carnemolla, B., Zardi, L., Neri, P. & Neri, D. (1998). Design and use of a phage display library-human antibodies with subnanomolar affinity against a marker of angiogenesis eluted from a two-dimensional gel. *J. Biol. Chem.* 273, 21769-21776.

Proba, K., Worn, A., Honegger, A. & Plückthun, A. (1998). Antibody scFv fragments without disulfide bonds made by molecular evolution. *J. Mol. Biol.* 275, 245-253.

Prodromou, C. & Pearl, L. H. (1992). Recursive PCR: a novel technique for total gene synthesis *Protein Eng.* 5, 827-829.

Rajewsky, K. (1996). Clonal selection and learning in the antibody system. *Nature*, 381, 751-758.

Roberts, R. J. & Macelis, D (1999). REBASE-restriction enzymes and methylases. *Nucl. Acids Res* 27, 312-313.

Rodi, D. J. & Makowski, L. (1999). Phage-display technology-finding a needle in a vast molecular haystack. *Curr. Opin. Biotechnol.* 10, 87-93.

Rodrigues, M. L., Presta, L. G., Kotts, C. E., Wirth, C., Mordenti, J., Osaka, G., Wong, W. L. T., Nuijens, A., Blackburn, B. & Carter, P. (1995). Development of a humanized disulfide-stabilized anti-beta P185(HER2) Fv-beta-lactamase fusion protein for activation of a cephalosporin doxorubicin prodrug. *Cancer Res.* 55, 63-70.

Rosok, M. J., Eghtedarzadehkondri, M., Young, K., Bajorath, J., Glaser, S. & Yelton, D. (1998). Analysis of Bl36 binding sites for antigen and anti-idiotype by codon-based scanning mutagenesis. *J. Immunol.* 160, 2353-2359.

Saul, F. A. & Poljak, R. J. (1993). Structural patterns at residue positions 9, 18, 67 and 82 in the VH framework regions of human and murine immunoglobulins. *J. Mol. Biol.* 230, 15-20.

Schier, R., Bye, J., Apell, G., McCall, A., Adams, G. P., Malmqvist, M., Weiner, L. M. & Marks, J. D. (1996a). Isolation of high-affinity monomeric human anti-c-erbb-2 single chain Fv using affinity-driven selection. *J. Mol. Biol.* 255, 28-43.

Schier, R., McCall, A., Adams, G. P., Marshall, K. W., Merritt, H., Yim, M., Crawford, R. S., Weiner, L. M., Marks, C. & Marks, J. D. (1996b). Isolation of picomolar affinity anti-c-erbb-2 single-chain Fv by molecular evolution of the complementarity determining regions in the center of the antibody binding site. *J. Mol. Biol.* 263, 551-567.

Searle, S. J., Pedersen, J. T., Henry, A. H., Webster, D. M. & Reos, A. R. (1995). Expressing antibodies in *Escherichia coli*. In *Antibody Engineering* (Borrebaeck, C. A. K., ed.), pp. 3-51, Oxford University Press, Oxford.

Sheets, M. D., Amersdorfer, P., Finnern, R., Sargent, P., Lindqvist, E., Schier, R., Hemingsen, G., Wong, C., Gerhart, J. C. & Marks, J. D. (1998). Efficient construction of a large nonimmune phage antibody library-the production of high-affinity human single-chain antibodies to protein antigens. *Proc. Natl Acad. Sci. USA*, 95, 6157-6162.

Shiloi, H., Kidera, A. & Nakamura, H. (1996). Structural classification of CDR-H3 in antibodies. *FEBS Letters*, 399, 1-8.

Smith, G. P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228, 1315-1317.

Smith, G. P. & Scott, J. K. (1993). Libraries of peptides and proteins displayed on filamentous phage. *Methods Enzymol.* 217, 228-257.

Söderlind, E., Vergeles, M. & Borrebaeck, C. A. K. (1995). Domain libraries: synthetic diversity for de novo design of antibody V-regions. *Gene*, 160, 269-272.

Sondek, J. & Shortle, D. (1992). A general strategy for random insertion and substitution mutagenesis: substoichiometric coupling of trinucleotide phosphoramidites. *Proc. Natl Acad. Sci. USA*, 89, 3581-3585.

Spada, S., Krebber, C. & Plückthun, A. (1997). Selectively infective phages (SIP). *Biol. Chem.* 378, 445-456.

Spada, S., Honegger, A. & Plückthun, A. (1998). Reproducing the natural evolution of protein structural features with the selectively infective phage (SIP) technology: the kink in the first strand of antibody kappa domains *J. Mol. Biol.* 283, 395-407.

Steipe, B., Schiller, B., Plückthun, A. & Steinbacher, S. (1994). Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* 240, 188-192.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22, 4673-4680.

Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). The repertoire of human germline VH sequences reveals about 50 groups of VH segments with different hypervariable loops. *J. Mol. Biol.* 227, 776-798.

Tomlinson, I. M., Cox, J. P., Gherardi, E., Lesk, A. M. & Chothia, C. (1995). The structural repertoire of the human V kappa domain. *EMBO J.* 14, 4628-4638.

Tomlinson, I. M., Walter, G., Jones, P. T., Dear, P. H., Sonnhammer, E. L. L. & Winter, G. (1996). The imprint of somatic hypermutation on the repertoire of human germline V genes. *J. Mol. Biol.* 256, 813-817.

Ulrich, H. D., Patten, P. A., Yang, P. L., Romesberg, F. E. & Schultz, P. G. (1995). Expression studies of catalytic antibodies. *Proc. Natl Acad. Sci. USA*, 92, 11907-11911

van Dijk, K. W., Mortari, F., Kirkham, P. M., Schroeder, H. W. & Milner, E. C. (1993). The human immunoglobulin VH7 gene family consists of a small, polymorphic group of six to eight gene segments dispersed throughout the VH locus. *Eur. J. Immunol.* 23, 832-839.

Vaughan, T. J., Williams, A. J., Pritchard, K., Osbourn, J. K., Pope, A. R., Earnshaw, J. C., McCafferty, J., Hodits, R. A., Wilton, J. & Johnson, K. S. (1996). Human antibodies with sub-nanomolar affinities isolated from a large non-immunized phage display library. *Nature Biotech.* 14, 309-314.

Vimekäs, B., Ge, L., Plückthun, A., Schneider, K. C., Wellnhofer, C. & Moroney, S. E. (1994). Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucl. Acids Res.* 22, 5600-5607.

Wagner, S. D. & Neuberger, M. S. (1996). Somatic hypermutation of immunoglobulin genes. *Annu. Rev. Immunol.* 14, 441-457.

Wall, J. G. & Plückthun, A. (1999). The hierarchy of mutations influencing the folding of antibody domains in *Escherichia coli*. *Protein Eng.* 12, 605-611.

Williams, S. C., Frippiat, J. P., Tomlinson, I. M., Ignatovich, O., Lefranc, M. P. & Winter, G. (1996). Sequence and evolution of the human germline V-lambda repertoire. *J. Mol. Biol.* 264, 220-232.

Winter, G. (1998). Synthetic human antibodies and a strategy for protein engineering. *FEBS Letters*, 430, 92-94.

Winter, G., Griffiths, A. D., Hawkins, R. E. & Hoogenboom, H. R. (1994). Making antibodies by phage display technology *Annu Rev. Immunol.* 12, 433-455.

Wörn, A. & Plückthun, A. (1998). An intrinsically stable antibody scFv fragment can tolerate the loss of both disulfide bonds and fold correctly. *FEBS Letters*, 427, 357-361.

Wörn, A. & Plückthun, A. (1999). Different equilibrium stability behavior of scFv fragments: Identification, classification and improvement by protein engineering. *Biochemistry*, 38, 8739-8750.

Wörn, A., Auf der Maur, A., Escher, D., Honegger, A., Barberis, A. & Plückthun, A. (2000). Correlation between in vitro stability and in vivo performance

of anti-GCN4 intrabodies as cytoplasmic inhibitors. *J. Biol. Chem.* in the press.

Wu, H., Beuerlein, G., Nie, Y., Smith, H., Lee, B. A., Hensler, M., Huse, W. D. & Watkins, J. D. (1998). Stepwise in vitro affinity maturation of Vitaxin, an alpha(v)beta(3)-specific humanized Mab. *Proc. Natl Acad. Sci. USA*, 95, 6037-6042.

Wu, T. T., Johnson, G. & Kabat, E. A. (1993). Length distribution of CDR1B in antibodies. *Proteins: Struct. Funct. Genet.* 16, 1-7.

Yang, W. P., Green, K., Pinz-Sweeney, S., Briones, A. T., Burton, D. R. & Barbas, C. F. (1995). CDR walking mutagenesis for the affinity maturation of a potent human anti-HIV-1 antibody into the picomolar range. *J. Mol. Biol.* 25, 392-403.

Yanisch-Perron, C., Vieira, J. & Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene*, 33, 103-119.

Yélamos, J., Klix, N., Goyenechea, B., Lozano, F., Chui, Y. L., González, Fernández A., Pannell, R., Neuberger, M. S. & Milstein, C. (1995). Targeting of non-Ig sequences in place of the V segment by somatic hypermutation. *Nature*, 376, 225-229.

Edited by I. A. Wilson

(Received 12 August 1999; received in revised form 3 December 1999; accepted 6 December 1999)

http://www.academicpress.com/jmb

Supplementary material for this paper is available from JMB Online.